

明 細 書

生物学的物質の配列情報の記録方法及び装置

5 関連する出願への言及

本願は、2002年4月17日に出願された国際出願でPCT第21条(2)のもとで英語では公開されていないPCT/JPO2/03801の一部継続出願である。本願及びその国際出願は、米国特許法第119条のもとで、2001年4月18日付け提出の日本国特願2001-120335及び2001年11月30日付け提出の日本国特願2001-368002に対して優先権を主張する。また、明細書、特許請求の範囲、図面、及び要約を含む2001年4月18日付け提出の日本国特願2001-120335、2001年11月30日付け提出の日本国特願2001-368002、及び2002年10月16日付け提出の米国特許出願第10/272,107号の全ての開示内容は、そっくりそのまま引用して本願に組み込まれている。

技術分野

本発明は、例えばDNA(デオキシリボ核酸:deoxyribonucleic acid)、RNA(リボ核酸:ribonucleic acid)、若しくは遺伝子等の核酸の少なくとも一部を構成する一列のヌクレオチド又はタンパク質の少なくとも一部を構成する一列のアミノ酸などの生物学的物質の配列情報の記録方法及び装置に関する。更に本発明は、その配列情報を供給するためのビジネスモデルとして好適な配列情報の供給方法、及びその配列情報を記録するコンピュータ読み取り可能な記録媒体に関する。

25

背景技術

人間、及び他の生物(動物、植物、微生物等)のDNAを構成する1対のヌクレオチドの鎖(又は塩基の鎖)の配列情報の解読が世界的に行われている。この場合、従来よりDNAを構成する4種類のヌクレオチドは、塩基としてアデニン

を含むヌクレオチド、グアニンを含むヌクレオチド、シトシンを含むヌクレオチド、及びチミンを含むヌクレオチドにそれぞれ文字A, G, C, 及びTを割り当てることによって、それぞれ1バイト (= 8ビット) のテキストデータで表わされている。その結果として一つのDNAの配列は、それを構成する1対の重合体の鎖の内の一方向の鎖のヌクレオチド (n個とする) の配列を順次文字A, G, C, T (又はa, g, c, t) の何れかで表すことによって、nバイトのテキストデータで表されていた。同様に、一つのRNAを構成する1本のn個のヌクレオチドの配列は、チミンを含むヌクレオチドの代わりにウラシルを含むヌクレオチドに文字U (又はu) を割り当てることによって、nバイトのテキストデータで表されていた。

これに関して、例えば人間の最も大きい第1染色体中のDNAの配列は、約2億5千万個のヌクレオチドの配列であり、最も小さい第22染色体中のDNAの配列は、約5000万個のヌクレオチドの配列であるため、人間の各染色体中のDNAの配列は、約250Mバイト~50Mバイトのテキストデータで表すことができる。更に、一人の人間の全部のDNA情報 (ゲノム) は、約30億個のヌクレオチドの配列で表すことができるため、そのゲノムは、約3Gバイトのテキストデータで記録することができる。なお、それらのテキストデータに対して通常のファイル圧縮技術を適用することによって、それらのテキストデータは、例えば元のデータの50%程度の圧縮ファイルとしても記録、又は送信することができる。

また、DNAの配列の解読に続いて、DNA中の多数の遺伝子の情報に基づいてそれぞれ合成されるタンパク質の機能の研究も広く行われている。この場合、タンパク質を構成する20種類のアミノ酸は、三文字表記 (3-Letter Code) ではそれぞれ3文字 (例えばAla, Cys, Glu等) のテキストデータで表され、一文字表記 (1-Letter Code) ではそれぞれ1文字のテキストデータ (例えばA, C, E等) で表されるため、n個のアミノ酸よりなるタンパク質の配列は、nバイトのテキストデータで表すことができる。そして、種々のタンパク質は、それらのアミノ酸が約20個~約1000個程度所定の順序で配列されたものであるため、それらのタンパク質の配列は、最大でも約1kバイト程度のテキストデー

タで記録することができる。また、例えば人間の遺伝子の総数は約3万個と言われており、タンパク質は理論的なものも含めて約10万種類の存在が可能であると言われている。

上記の如く例えば一人の人間のDNA情報をテキストデータで記録するためには、全部で3Gバイト程度の記憶容量が必要であり、仮に通常の圧縮ファイルの技術を適用しても1Gバイト程度の記憶容量が必要である。また、人間以外の大腸菌や各種ウィルス等のDNA情報も解析されて次第に公開されるようになっていくが、これらのDNA情報をテキストデータの形で多く集めると、数100Gバイト程度の記憶容量が必要である。これはRNAの配列情報についても同様である。

このように人間又は他の生物のDNA情報をテキストデータ、又はこの通常の圧縮ファイルの形で記録するものとする、例えば1枚の記憶容量が5Gバイト程度のDVD-ROM (digital video disc-ROM) ディスクのように膨大な記憶容量を持つ記録媒体が必要である。更に、そのDNA情報を利用する場合にその記録媒体からの読み出し時間が長くなり、処理時間が長くなるという不都合がある。

また、現状の一般の通信回線の通信速度は、最大で5Mbps程度であるため、例えば1Gバイト程度のDNA情報をその通信回線を介して送信するものとする、送信時間は最短でも約30分程度となる。特に最近はそのDNA情報をデジタルの携帯電話システムを介して送信する場合も考えられるが、現在の携帯電話システムの通信速度はせいぜい1Mbps程度であるため、少なくとも人間のDNA情報の伝送で使用することは現状ではあまり実用的ではない。

次に、例えば或る微生物のDNA中の遺伝子について複数の研究者が並行して研究するような場合に、複数の研究者が保有している標準となるDNAのヌクレオチドの配列の同一性をどのように保証するのかという問題がある。即ち、そのDNAのヌクレオチドの配列が例えば数Mバイト（文字数で数100万文字）程度のテキストデータで記録されている場合に、複数の研究者が互いに自分のテキストデータと他人のテキストデータとの同一性（完全一致性）を短時間に確認するのは必ずしも容易ではない。

これに関連して、例えば人間又は他の生物のDNA情報の利用方法としては、

標準的なDNAの配列と、検査対象のDNAの配列との間の相違する部分をサーチする場合が考えられる。これは、いわゆるSNP（一塩基変位多型：Single Nucleotide Polymorphism）の可能性を検査するような場合に必要になると考えられる。しかしながら、両方のDNAのヌクレオチドの配列がそれぞれ膨大なテキストデータで表わされている場合に、それら2つのテキストデータを比較して相違点を検出するにはかなりの長い時間が必要となり、検査時間が長くなるという不都合がある。

更に、人間又は他の生物のDNA情報を製薬会社の研究者等のユーザに提供するビジネスも行われつつあるが、この場合に、情報供給者が例えば通信回線を介してDNA情報をユーザに提供する場合には、できるだけ少ない情報量で、即ち短い送信時間で必要な情報をユーザに提供できるビジネスモデルが必要である。また、ユーザ側では、提供されたDNA情報に伝送エラー等が無いかどうかを容易に確認できることが望ましい。上記の各課題はRNAのヌクレオチドの配列情報についても同様に当てはまるものである。

更に、一つのタンパク質のアミノ酸の配列は、最大でも約1kバイト程度のテキストデータで記録することができるが、タンパク質の種類は理論的に約10万个程度にもなるため、全部のタンパク質の配列情報をテキストデータで表すと、全部のDNAの配列情報程度の膨大な量となる。従って、個々のタンパク質の配列は、できるだけ少ない情報量で記録できることが望ましい。また、2つのタンパク質の配列情報の相違部を容易に確認できるシステムも必要である。

本発明は斯かる点に鑑み、核酸中の一系列のヌクレオチド、又はタンパク質中の一系列のアミノ酸などの生物学的物質の配列情報を近似的に少ないデータ量で記録できる記録方法及び装置を提供することを第1の目的とする。

また、本発明は、2つの生物学的物質の配列情報同士の相違する部分を少ないデータ量で容易に検出できると共に、必要に応じてその相違する部分の情報を復元できる記録方法及び装置を提供することを第2の目的とする。

また、本発明は、一系列のヌクレオチド、又は一系列のアミノ酸などの生物学的物質の配列情報をユーザに提供する場合に、ユーザが提供された配列情報と情報供給者が保持している配列情報との相違する部分を少ないデータ量で容易に確認で

きるビジネスモデル（情報供給方法）を提供することを第3の目的とする。

また、本発明は、生物学的物質の配列情報が少ないデータ量で近似的に記録されたコンピュータ読み取り可能な記録媒体を提供することを第4の目的とする。

5 発明の開示

本発明による第1の生物学的物質の配列情報の記録方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが m ビット（ m は16以上の整数）の部分データ（ $A(i, j)$ ）に分割し、複数行のその部分データに各行毎にその非配列方向にガロア体 $GF(2^m)$ 上の第1の演算を施して第1組のパリティ情報（ $B1(i)$, $B2(i)$, $B3(i)$ ）を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体 $GF(2^m)$ 上の第2の演算を施して第2組のパリティ情報（ $C1(j)$, $C2(j)$, $C3(j)$ ）を求め、その第1組及び第2組のパリティ情報でその生物学的物質の配列を表すものである。

斯かる本発明によれば、その生物学的物質としては、例えば一列のヌクレオチド又は一列のアミノ酸が考えられる。前者の一列のヌクレオチドは、例えば或るDNA (deoxyribonucleic acid) を構成する1対の重合体の鎖の一方の鎖の少なくとも一部、或るRNA (ribonucleic acid) を構成する1列の重合体の鎖の少なくとも一部、又は或る遺伝子の少なくとも一部である。その一列のヌクレオチドの配列は、各ヌクレオチドに含まれる塩基の配列ともみなすことができる。一方、後者の一列のアミノ酸は、例えば或るタンパク質を構成するアミノ酸の配列の少なくとも一部である。

その生物学的物質の全体の個数を NT として、各生物学的物質をそれぞれ1文字（例えばアルファベット）で表すものとする、その生物学的物質の配列に対応するテキストデータの全体の量は、例えばアスキーコード (ASCII code) (ANSI形式) では NT バイトになり、ユニコード (Unicode) では $2 \cdot NT$ バイトになる。なお、配列を見易くするためのスペース、数字、及び改行などのコード

は無視している。そして、例えば図7の例において、テキストデータを配列方向に N 個 ($i = 1 \sim N$) で、非配列方向に M 個 ($j = 1 \sim M$) の部分テキストデータ $T(i, j)$ に分割し、図8に示すように、各部分テキストデータ $T(i, j)$ をそれぞれ m ビットの部分データ $A(i, j)$ に変換する。 m ビットの部分データ $A(i, j)$ は、それぞれ n 個 (図8の例では $n = 16$) の連続する生物学的物質の配列を表している。

この場合、最も簡単な方法としては、部分データ $A(i, j)$ として部分テキストデータ $T(i, j)$ そのものを数値データとみなしたデータを使用すればよい。即ち、テキストデータがアスキーコードで記録されている場合には、部分データ $A(i, j)$ としてはそのアスキーコードを使用すればよい。また、テキストデータがユニコードで記録されている場合には、各文字をそれぞれの2バイトのコードの上位1バイトで表したものを部分データ $A(i, j)$ としてもよい。但し、処理対象のデータ量を少なくするためには、各生物学的物質を表す文字を例えば6ビット以下の数値データに変換する変換テーブル (所定の規則) を用いて、部分テキストデータ $T(i, j)$ を変換して得られる数値データを部分データ $A(i, j)$ とすることが望ましい。

次に、 m ビットの各部分データ $A(i, j)$ を非配列方向、及び配列方向に演算することによって、各行及び各列の配列情報を近似的に表すデータを算出する。このためには、 m ビットのデータを加減乗除の対象にできる体 (Field) が必要であり、本発明ではそのために第1の方法としてガロア体 (拡大ガロア体) $GF(2^m)$ を用いる。ガロア体 $GF(2^m)$ を用いた場合には、各行又は各列毎に m ビットの部分データ $A(i, j)$ 、及び必要に応じて m ビットの係数を用いて所定の加減乗除演算 (第1又は第2の演算) を行ったときに得られる一つの情報 (これを本発明では「パリティ情報」と呼ぶ。) が m ビットであるため、配列情報を少ないデータ量で簡潔に記録できる利点がある。

2を法とする数 (0及び1) で表される体を Z_2 とすると、ガロア体 $GF(2^m)$ 上の演算は、体 Z_2 上の係数を持つ m 次の既約多項式 $GF(X)$ を用いて定義することができる。即ち、2つの m ビットの部分データ $A(i, j)$ 、 $A(i', j')$ をそれぞれ2進数表示で $(a_{m-1} \ a_{m-2} \ \cdots \ a_1 \ a_0)$ 、 $(b_{m-1} \ b$

$a_{m-2} \cdots b_1 b_0$) とすると (a_k, b_k は0又は1)、これらをそれぞれ次のように $(m-1)$ 次以下の多項式 $A F(X)$, $B F(X)$ に変換する。

$$A F(X) = a_{m-1} \cdot X^{m-1} + a_{m-2} \cdot X^{m-2} + \cdots + a_1 \cdot X + a_0 \cdots (1)$$

$$B F(X) = b_{m-1} \cdot X^{m-1} + b_{m-2} \cdot X^{m-2} + \cdots + b_1 \cdot X + b_0 \cdots (2)$$

- 5 この場合、ガロア体 $GF(2^m)$ 上で $A F(X)$ と $B F(X)$ とを加算する場合には、 X の各次数 k ($k=0 \sim (m-1)$) において、係数 a_k と係数 b_k とを体 Z_2 上で加算すればよい。体 Z_2 上では加算と減算とは同じ結果になる。この結果、得られた多項式の係数を2進数表示で表したものの(ベクトル表示)が、部分データ $A(i, j)$, $A(i', j')$ をガロア体 $GF(2^m)$ 上で加算した結果になる。これは、ビット毎に排他的論理和演算を行うのと同じ結果である。

- 10 次に、ガロア体 $GF(2^m)$ 上で $A F(X)$ に $B F(X)$ を乗算する場合には、先ず通常の乗算を行って積を求めた後、この積を既約多項式 $G F(X)$ で除算した後の余りの多項式 $C F(X)$ を次のように求める (c_k は0又は1)。これを既約多項式 $G F(X)$ を法(modulus)とする乗算と呼ぶ。この際にも X の各次数での係数の加算(減算)は体 Z_2 上で行われる。

$$C F(X) = c_{m-1} \cdot X^{m-1} + c_{m-2} \cdot X^{m-2} + \cdots + c_1 \cdot X + c_0 \cdots (3)$$

- この多項式 $C F(X)$ の係数を2進数表示で表したものの($c_{m-1} c_{m-2} \cdots c_1 c_0$) が、部分データ $A(i, j)$, $A(i', j')$ をガロア体 $GF(2^m)$ 上で乗算した結果になる。また、任意の m ビットの係数を β とすると、係数 β も(2)式と同様の $(m-1)$ 次以下の多項式 $D F(X)$ で表される。従って、例えば部分データ $A(i, j)$ に係数 β を乗算する場合には、(1)式の多項式 $A F(X)$ と多項式 $D F(X)$ との積を既約多項式 $G F(X)$ を法として計算すればよい。また、例えば部分データ $A(i, j)$ を係数 β で除算する場合には、部分データ $A(i, j)$ に β の逆元 β^{-1} を乗算すればよい。

- 25 従って、 m ビットの全てのデータ(全ての部分データ $A(i, j)$ が含まれる)は、ガロア体 $GF(2^m)$ 上のベクトル表示での元とみなすことができ、 m ビットのデータは、多項式表示では、(1)式のような $(m-1)$ 次以下の多項式で表すことができる。また、生物学的物質の配列(文字列)を部分データ $A(i, j)$ に対応させる変換テーブル(所定の規則)の逆変換を用いて、必要に

応じてそのベクトル表示のmビットのデータを文字列に変換することによって、そのデータに対応する生物学的物質の配列が得られる。

そして、本発明では、例えば図8に示すように、部分データA (i, j) が配列方向にN個 (i = 1 ~ N) で、非配列方向にM個 (j = 1 ~ M) で配列され、
5 各行毎に第1組のパリティ情報 (B 1 (i), B 2 (i), B 3 (i)) が得られ、各列毎に第2組のパリティ情報 (C 1 (j), C 2 (j), C 3 (j)) が得られる。これら2組のパリティ情報の内の1つのパリティ情報 (例えばB 1 (1)) はそれぞれ1つの部分データA (i, j) と同じmビットのデータで表される。

10 この場合の部分データA (i, j) の全体のデータ量DT 1は、以下のようになる。

$$DT\ 1 = m \cdot N \cdot M \text{ (ビット)} \quad \cdots (4)$$

また、第1組及び第2組のパリティ情報が、それぞれe個 (eは1以上の整数) のパリティ情報を含むとすると、パリティ情報全体のデータ量DS 1は、
15 下のようになる。なお、e個のパリティ情報を含む場合には、各行及び各列において、それぞれe個までの部分データA (i, j) を復元できる。

$$DS\ 1 = m \cdot e \cdot (N + M) \text{ (ビット)} \quad \cdots (5)$$

従って、例えば生物学的物質がDNAを構成するヌクレオチドであるとして、仮にN = 64, M = 128, e = 2とすると、(4)式及び(5)式よりデータ
20 量DT 1, DS 1は以下のようになる。

$$DT\ 1 = m \cdot 8192 \text{ (ビット)} \quad \cdots (6)$$

$$DS\ 1 = m \cdot 384 \text{ (ビット)} \cong DT\ 1 / 20 \quad \cdots (7)$$

従って、パリティ情報のデータ量は、部分データA (i, j) 全体のデータ量のほぼ1/20程度に少なくできる。この場合、例えば人間の1本の染色体のDNAの配列は、50Mバイト~250Mバイト程度のテキストデータで表されるため、予めそのテキストデータを500個~2500個程度のブロックに分割し、
25 各ブロック毎に2組のパリティ情報を求めることによって、全部のパリティ情報のデータ量はそのテキストデータのほぼ1/20程度、即ち2.5Mバイト~12.5Mバイト程度に少なくできる。また、その部分データA (i, j) が、例

例えばテキストデータを $1/f$ に小さくしたデータ量である場合には、パリティ情報も更に $1/f$ だけ少なくすることができる。

本発明によれば、元の生物学的物質の配列情報を近似的に表す情報（パリティ情報）を、元のテキストデータよりも少ないデータ量のファイルに記録することができる。従って、記録媒体として、DVD-ROMのような大容量の媒体の他に、CD-ROM、及びフラッシュROMのような小容量でも通常のコンピュータで手軽に再生できる媒体を使用できる。更に、少ないデータ量の配列情報であれば、通信回線を介して短時間に送信できるため、そのパリティ情報は、例えば携帯電話システムなどを介してユーザに安価に供給することも可能となる。

そして、第1組のパリティ情報、及び第2組のパリティ情報を用いることによって、ユーザ側では、2つの生物学的物質の配列の相違する部分を容易に特定することができると共に、相違する部分の個数が各行、又は各列で e 個以下である場合には、パリティ情報を用いて相違する部分の配列の復元を行うことも可能となる。

なお、テキストデータが記録されたファイルが通常の圧縮技術（ZIPファイル、LHAファイル等）で圧縮できるように、本発明のパリティ情報が記録されたファイルも更に通常の圧縮技術を用いて圧縮して記録できることは言うまでもない。但し、圧縮されたファイルを使用する場合には、解凍作業が必要になり、最終的には元のファイルを復元する必要があるため、元のファイル自体のデータ量を減らしておくことは極めて有効である。

次に、上記の本発明において、そのガロア体 $GF(2^m)$ 上の生成元を α としたとき、一例として、その第1組のパリティ情報は、複数行の各行のその部分データ $(A(i, j))$ にそれぞれその非配列方向に順次 $\alpha^{s \cdot p}$, $\alpha^{s \cdot (p+1)}$, $\alpha^{s \cdot (p+2)}$, ..., $\alpha^{s \cdot (p+d_p)}$ (s は0以上の整数、 p は0以上の整数、 d_p は1以上の整数) を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第2組のパリティ情報は、複数列の各列のその部分データ $(A(i, j))$ にそれぞれその配列方向に順次 $\alpha^{t \cdot q}$, $\alpha^{t \cdot (q+1)}$, $\alpha^{t \cdot (q+2)}$, ..., $\alpha^{t \cdot (q+d_q)}$ (t は0以上の整数、 q は0以上の整数、 d_q は1以上の整数) を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。

この場合、 $p = q = 0$ とすると、第1組のパリティ情報 $B_1(i)$ 、及び第2組のパリティ情報 $C_1(j)$ は、それぞれガロア体 $GF(2^m)$ 上の次の演算によって計算される。(8)式の Σ は j について $1 \sim M$ までの和を表し、(9)式の Σ は i について $1 \sim N$ までの和を表している。

$$5 \quad B_1(i) = \Sigma \alpha^{s(j-1)} \cdot A(i, j) = A(i, 1) + \alpha^s \cdot A(i, 2) \\ + \alpha^{2s} \cdot A(i, 3) + \dots + \alpha^{(M-1)s} \cdot A(i, M) \quad \dots (8)$$

$$C_1(j) = \Sigma \alpha^{t(i-1)} \cdot A(i, j) = A(1, j) + \alpha^t \cdot A(2, j) \\ + \alpha^{2t} \cdot A(3, j) + \dots + \alpha^{(N-1)t} \cdot A(N, j) \quad \dots (9)$$

そして、(8)式、(9)式で $s = t = 0$ とすると、パリティ情報 $B_1(i)$ 、
 10 $C_1(j)$ は、それぞれ部分データ $A(i, j)$ のガロア体 $GF(2^m)$ 上の和、即ち各行又は各列で部分データ $A(i, j)$ に排他的論理和演算を施して得られる値を示す。従って、簡単な演算で、各行及び各列の配列の近似的な情報を求めることができる。但し、この場合には、各行又は各列で2つの部分データ $A(i, j)$ が入れ替わったような配列に対しても、パリティ情報 $B_1(i)$ 、
 15 $C_1(j)$ は同じ値になってしまう。

次に、 $s = t = 1$ とすると、パリティ情報 $B_1(i)$ 、 $C_1(j)$ は、それぞれ各行又は各列で部分データ $A(i, j)$ に $1, \alpha, \alpha^2, \alpha^3, \dots$ を乗算して得られる積の和を示す。この場合、各行又は各列で2つの部分データ $A(i, j)$ が入れ替わったような配列に対しても、パリティ情報 $B_1(i)$ 、
 20 $C_1(j)$ は異なった値となるため、例えば2つの生物学的物質の配列間の相違する部分をより正確に特定することができる。そして、或る $s (\neq 0)$ (又は $t (\neq 0)$) の値において、そのように各行又は各列で2つの部分データに常に異なる係数を乗ずるためには、係数 $\alpha^{s(j-1)}$ (又は $\alpha^{t(i-1)}$) は互いに異なる必要がある。そのためには、 α をガロア体 $GF(2^m)$ 上の生成元として、各行及び各列
 25 の部分データ $A(i, j)$ の個数を $(2^m - 1) / s$ (又は $(2^m - 1) / t$) 以下とすればよい。即ち、 α を生成元とすることによって、扱う生物学的物質の配列を最も大きくすることができる。

また、各行及び各列において、それぞれ1つのパリティ情報を用いることによって、2つの生物学的物質の配列を比較する場合に、各行及び各列における1つ

の部分データ $A(i, j)$ の相違部を正確に復元することができる。従って、例えば遺伝子中の一つの塩基（ヌクレオチド）だけが異なる SNP（一塩基変位多型：Single Nucleotide Polymorphism）は本発明によって容易に検出できると共に、それに対応する正常な配列も容易に復元できる。

- 5 更に、各行及び各列における複数個 s' 及び t' の部分データ $A(i, j)$ の相違部を正確に復元するためには、その第 1 組のパリティ情報 ($B_1(i), B_2(i), B_3(i)$) は、その複数行の各行毎にその整数 s について互いに異なる複数 (s' 個) の値で求めた複数の和を含み、その第 2 組のパリティ情報 ($C_1(j), C_2(j), C_3(j)$) は、その複数列の各列毎にその整数 t
- 10 について互いに異なる複数 (t' 個) の値で求めた複数の和を含むようにすればよい。その相違部を復元するためには、ガロア体 $GF(2^m)$ 上で s' 元 (t' 元) の 1 次連立方程式を解けばよい。

- また、本発明においては、その部分データのその配列方向の個数を、その部分データのその非配列方向の個数よりも少なくして、その第 2 組のパリティ情報の
- 15 個数を、その第 1 組のパリティ情報の個数よりも少なくしてもよい。

- 特に本発明によって得られるパリティ情報をディスプレイに表示するような場合には、その部分データの配列方向の個数をそのディスプレイの横幅に対応する数に制限し、その部分データの非配列方向の個数を大きくすることによって、その非配列方向のデータは、その表示画面上での上下方向へのスクロールによって
- 20 容易に表示できるため、配列情報を効率的に、且つ分かり易く表示できる。

- 但し、このように部分データの配列方向の個数が非配列方向の個数よりも少ないときには、その第 1 組及び第 2 組のパリティ情報を同じ量にすると、パリティ情報の情報量が全体として多くなる。そこで、配列方向のパリティ情報（第 2 組のパリティ情報）の個数を非配列方向のパリティ情報（第 1 組のパリティ情報）
- 25 の個数よりも少なくすることによって、パリティ情報を少なくして、且つ 2 つの配列間の相違部の復元を効率的に行うことができる。

また、その部分データの前記非配列方向の個数を、 $(2^m - 1) / 4$ 以下にすることが望ましい。これによって、その非配列方向において 4 つの異なる係数 ($\alpha^k, \alpha^{2k}, \alpha^{3k}, \alpha^{4k}$) を乗ずることができるため、2 つの配列間の各行の

相違部（非配列方向の相違部）を4個まで正確に復元することができる。これは通常のSNPの検出などには十分であると思われる。

また、本発明において、そのガロア体 $GF(2^m)$ を規定する整数 m は64の倍数であることが望ましい。最近のコンピュータにはデータの処理単位が64ビットであるタイプが増加しているため、整数 m を64の倍数とすることによって、
5 効率的にパリティ情報を算出することができる。

また、本発明において、その生物学的物質の配列を基準配列として、この基準配列のその2組のパリティ情報に対応させて、検査対象の生物学的物質の配列についてその2組のパリティ情報を求め、その4組のパリティ情報よりその基準配列に対するその検査対象の生物学的物質の配列の相違部を求めるようにしてもよい。このように基準配列の2組のパリティ情報と、検査対象の2組のパリティ情報とを比較するのみで、相違部の位置を容易に検出できると共に、相違部が各行及び各列で所定個数以下であれば、4組のパリティ情報と検査対象の生物学的物質の配列とから、連立方程式を解くことによって、基準配列の相違部のデータを
10 正確に復元することもできる。
15

次に、本発明の生物学的物質の配列情報の記録装置は、その生物学的物質の配列情報を読み取る配列読み取り装置（4）と、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向
20 に交差する非配列方向に複数列の長さが m ビット（ m は16以上の整数）の部分データ（ $A(i, j)$ ）に分割するデータ配列手段（10、ステップ105）と、複数行のその部分データに各行毎にその非配列方向にガロア体 $GF(2^m)$ 上の第1の演算を施して第1組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体 $GF(2^m)$ 上の第2の演算を施して第
25 2組のパリティ情報を求める演算手段（10、ステップ106）と、その第1組及び第2組のパリティ情報を記録媒体に記録する記録手段（15）とを有するものである。これによって、本発明のヌクレオチドやアミノ酸などの生物学的物質の配列情報の記録方法が実施できる。

この本発明の記録装置において、そのガロア体 $GF(2^m)$ 上の生成元を α と

したとき、一例としてその第1組のパリティ情報は、複数行の各行のその部分データ ($A(i, j)$) にそれぞれその非配列方向に順次 $\alpha^{s \cdot p}$, $\alpha^{s \cdot (p+1)}$, $\alpha^{s \cdot (p+2)}$, ..., $\alpha^{s \cdot (p+d \cdot p)}$ (s は0以上の整数、 p は0以上の整数、 $d \cdot p$ は1以上の整数) を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第2組のパリティ情報は、複数列の各列のその部分データにそれぞれその配列方向に順次 $\alpha^{t \cdot q}$, $\alpha^{t \cdot (q+1)}$, $\alpha^{t \cdot (q+2)}$, ..., $\alpha^{t \cdot (q+d \cdot q)}$ (t は0以上の整数、 q は0以上の整数、 $d \cdot q$ は1以上の整数) を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。この場合、部分データ ($A(i, j)$) に乗ずる係数は、生成元 α だけから計算できるため、演算が単純化される。

また、本発明の記録媒体は、生物学的物質の配列情報を記録したコンピュータ読み取り可能な記録媒体 (16) であって、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが m ビット (m は16以上の整数) の部分データに分割し、複数行のその部分データに各行毎にその非配列方向にガロア体 $GF(2^m)$ 上の第1の演算を施して第1組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体 $GF(2^m)$ 上の第2の演算を施して第2組のパリティ情報を求め、その生物学的物質の配列に関する情報を、その第1組及び第2組のパリティ情報として記録したものである。

本発明によれば、ヌクレオチドやアミノ酸などの生物学的物質の配列情報を近似的に表すパリティ情報を少ないデータ量でその記録媒体に記録できるため、記録媒体としてCD-ROM, CD-R、フラッシュROMなどの記録容量は比較的少ないが、使い勝手の良い媒体をも使用できる。

この場合、その生物学的物質の配列に対応するそのテキストデータ、又はこのテキストデータに対応するその数値データの40ビット以上の長さの数学的な要約値 (message digest) を更にその記録媒体に記録することが望ましい。

その数学的な要約値は、その生物学的物質の配列に対応するテキストデータ又は数値データに例えばMD5ハッシュ関数 (要約値は128ビット)、又はSH

S (Secure Hash Standard) ハッシュ関数 (要約値は 1 6 0 ビット) などのハッシュ関数の演算を施して得られるものである。その要約値を用いることによって、比較対象の 2 つの生物学的物質の膨大な配列が一致するかどうかを高い確率で極めて容易に確認することができる。また、パリティ情報を用いて相違する部分データの復元を行った後に、要約値を比較することによって、データが完全に復元できたかを確認することができる。その要約値が 4 0 ビット以上であれば、例えば全人類の DNA のヌクレオチドの配列情報をほぼ互いに衝突することなく表すことができる。

この場合、ガロア体 $GF(2^m)$ を決定する整数 m が 6 4 の倍数であるときには、要約値が 6 4 ビットの倍数となるハッシュ関数 (例えば MD 5 ハッシュ関数) を用いることが望ましい。演算を効率的に実行できるからである。

次に、本発明の第 1 の生物学的物質の配列情報の供給方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを保持する供給者 (2 A) が、そのテキストデータ、又はこれに対応するその数値データを第 1 ファイル (1 9) に記録して保持する第 1 ステップ (ステップ 1 0 4) と、その第 1 ファイルに記録されているそのテキストデータ、又はこのテキストデータに対応するその数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが m ビット (m は 1 6 以上の整数) の部分データに分割し、複数行のその部分データに各行毎にその非配列方向にガロア体 $GF(2^m)$ 上の第 1 の演算を施して第 1 組のパリティ情報 ($B_1(i) \sim B_3(i)$) を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体 $GF(2^m)$ 上の第 2 の演算を施して第 2 組のパリティ情報 ($C_1(j) \sim C_3(j)$) を求める第 2 ステップ (ステップ 1 0 5, 1 0 6) と、その供給者が、その第 1 組及び第 2 組のパリティ情報を第 2 ファイル (2 0) に記録して保持する第 3 ステップ (ステップ 1 0 7) と、その生物学的物質の配列情報のユーザ (2 B) が、通信回線 (1) を介してその供給者よりその第 2 ファイルに記録されているその 2 組のパリティ情報を受け取る第 4 ステップ (ステップ 1 1 0, 1 2 9) とを有するものである。

この供給方法は、上記の本発明の生物学的物質の配列情報の記録方法を、その配列情報を供給（販売）する際のビジネスモデルに適用したものである。即ち、本発明のビジネスモデルでは、或る生物XのDNAのヌクレオチド、又はタンパク質のアミノ酸などの生物学的物質の配列を最初に解読した供給者は、その配列のテキストデータ（又はこれを変換した数値データ）より、その配列情報を少ないデータ量で近似するパリティ情報を算出し、これをその通信回線を介してユーザに供給する。上述のように、一例としてパリティ情報は、元のテキストデータの1/20程度のデータ量であるため、そのパリティ情報はその通信回線を介して短時間で受信することができる。

本発明の供給方法においては、更にそのユーザが、その2組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内のその供給者の生物学的物質の配列情報との相違部を特定する第5ステップ（ステップ130, 131）と、この相違部の配列の復元ができない場合に、そのユーザがその通信回線を介してその供給者よりその第1ファイルに記録されているそのテキストデータ、又はその数値データの内のその配列の復元ができない部分の配列情報を受け取る第6ステップ（ステップ135）とを有することが望ましい。

このようにユーザ側で、パリティ情報だけで検査対象の配列内の供給者の配列との相違部の特定、及び復元ができる場合には、それ以上の配列情報を購入する必要が無い。一方、相違部が多く存在し、パリティ情報のみでは全部の正確なデータが復元できない場合には、例えば復元できない部分のテキストデータ（又は数値データ）のみを購入することによって、通信回線を介して必要な配列情報を短時間に購入できる。従って、通信回線として、携帯電話システムのような比較的低速の通信回線も使用できる。

また、本発明の供給方法においては、そのガロア体GF(2^m)上の生成元を α としたとき、一例として、その第1組のパリティ情報は、複数行の各行のその部分データにそれぞれその非配列方向に順次 $\alpha^{s \cdot p}$, $\alpha^{s \cdot (p+1)}$, $\alpha^{s \cdot (p+2)}$, ..., $\alpha^{s \cdot (p+d \cdot p)}$ （sは0以上の整数、pは0以上の整数、dpは1以上の整数）を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第2組のパリティ情報は、複数列の各列のその部分データにそれぞれその配列方

向に順次 $\alpha^{l(q)}$, $\alpha^{l(q+1)}$, $\alpha^{l(q+2)}$, ..., $\alpha^{l(q+dq)}$ (t は 0 以上の整数、 q は 0 以上の整数、 dq は 1 以上の整数) を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。

これらのパリティ情報を用いることによって、そのユーザは、SNP (一塩基変位多型) などを容易に検出することができる。

また、その供給方法においては、その供給者は、その生物学的物質の配列の長さの情報、及びその配列を表すテキストデータ又はその数値データの数学的な要約値の情報をその通信回線を介して閲覧可能な状態にしておき、そのユーザは、その第 4 ステップの前にその通信回線を介してその配列の長さの情報及びその数

10 学的な要約値の情報を閲覧する (ステップ 1 2 1) ことが望ましい。

この場合、その供給者は、その生物 X の生物学的物質の配列のテキストデータ (又はこれを変換した数値データ) よりハッシュ関数によって算出した要約値 (message digest) を例えばインターネット上で閲覧可能にする。これによって、その供給者は、そのテキストデータ自体を公開することなく、最初にその生物 X

15 の生物学的物質の配列を解読したことを主張できる。更に、ユーザが同じ配列情報を異なる供給者から誤って購入することも防止できる。

また、或るユーザが、その供給者よりその生物学的物質の配列情報を購入した後、購入した配列情報よりそのハッシュ関数によって要約値を算出し、その配列の長さも求める。そして、この配列の長さ、及び要約値をインターネット上で公

20 開されている値と比較することによって、購入した配列情報が正確なものであるかどうかを極めて高い確率で確認できる。

この場合、一例として、その数学的な要約値は、40 ビット以上で 192 ビット以下のデータであり、その供給者は、更にその生物学的物質の所定の一部の配列の情報をその通信回線を介して閲覧可能な状態にしておくことが望ましい。その要約値、及びその配列の長さの他に、そのように例えばその配列の先頭の 8 個

25 程度、及び後端の 8 個程度の配列を比較することによって、同一性の確認をより高精度に行うことができる。

次に、本発明による第 2 の生物学的物質の配列情報の記録方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則

に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが m ビット（ m は16以上の整数）の部分データ（ $A(i, j)$ ）に分割し、その部分データの最大値を N_{max} 、この最大値 N_{max} よりも大きい素数を P として、複数行のその部分データに各行毎にその非配列方向にガロア体 $GF(P)$ 上の第1の演算を施して第1組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体 $GF(P)$ 上の第2の演算を施して第2組のパリティ情報を求め、その第1組及び第2組のパリティ情報でその生物学的物質の配列を表すものである。

斯かる本発明によれば、その生物学的物質としては、例えば一列のヌクレオチド又は一列のアミノ酸が考えられる。前者の一列のヌクレオチドは、例えば或るDNAの一方の鎖の少なくとも一部、或るRNAを構成する1列の重合体の鎖の少なくとも一部、又は或る遺伝子の少なくとも一部である。そして、例えば図7の例において、その生物学的物質の配列情報を示すテキストデータを配列方向に N 個（ $i = 1 \sim N$ ）で、非配列方向に M 個（ $j = 1 \sim M$ ）の部分テキストデータ $T(i, j)$ に分割し、図8に示すように、各部分テキストデータ $T(i, j)$ をそれぞれ m ビットの部分データ $A(i, j)$ に変換する。 m ビットの部分データ $A(i, j)$ は、それぞれ n 個（図8の例では $n = 16$ ）の連続する生物学的物質の配列を表している。

次に、 m ビットの各部分データ $A(i, j)$ を非配列方向、及び配列方向に演算することによって、各行及び各列の配列情報を近似的に表すデータを算出する。このためには、 m ビットのデータを加減乗除の対象にできる体（Field）が必要であり、本発明ではそのために第2の方法としてガロア体 $GF(P)$ を用いる。ガロア体 $GF(P)$ を用いた場合、その部分データ $A(i, j)$ の最大値が $(2^m - 1)$ であると、その素数 P は $(m + 1)$ ビットの値となり、各行及び各列の部分データにガロア体 $GF(P)$ 上の所定の演算を施して得られる一つの情報（これを本発明では「パリティ情報」と呼ぶ。）も $(m + 1)$ ビットで表されるため、各パリティ情報のデータ量が1ビットだけ大きくなる。しかしながら、全体として見ると、ガロア体 $GF(2^m)$ を用いる場合とほぼ同程度に少ないパリティ情

報で、元の配列情報を近似することができる。更に本発明によれば、そのパリティ情報を求めるための演算はガロア体 $GF(2^m)$ での演算に比べると簡単である。

5 本発明においても、2つの配列の比較を行う際にそのパリティ情報を比較することによって、相違部の特定、及び相違部の或る程度の復元を行うことができる。

本発明においては、その部分データの最大値 N_{max} は $(2^m - 1)$ よりも小さい値であることが望ましい。これを実現する最も簡単な方法は、部分データ $A(i, j)$ として、この部分データに対応する配列を表すテキストデータそのもの（数値データとみなす）を使用することである。このとき、その素数 P は、次
10 の関係を満たすように選択できることが望ましい。

$$2^m > P > N_{max} \quad \cdots (A1)$$

これは、その素数 P をその最大値 N_{max} よりも大きい m ビットの数とすることを意味する。これによって、パリティ情報も m ビットのデータとなるため、ガロア体 $GF(2^m)$ よりも簡単な演算でよいことと、ガロア体 $GF(2^m)$ を用
15 いる場合と同じデータ量でパリティ情報を表すことを両立できる。

この場合、そのガロア体 $GF(P)$ 上の生成元を δ としたとき、一例としてその第1組のパリティ情報は、複数行の各行のその部分データにそれぞれその非配列方向に順次 $\delta^{s \cdot p}$, $\delta^{s \cdot (p+1)}$, $\delta^{s \cdot (p+2)}$, ..., $\delta^{s \cdot (p+d_p)}$ (s は0以上の整数、 p は0以上の整数、 d_p は1以上の整数) を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第2組のパリティ情報は、複数
20 列の各列のその部分データにそれぞれその配列方向に順次 $\delta^{t \cdot q}$, $\delta^{t \cdot (q+1)}$, $\delta^{t \cdot (q+2)}$, ..., $\delta^{t \cdot (q+d_q)}$ (t は0以上の整数、 q は0以上の整数、 d_q は1以上の整数) を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。その整数 s , t の値を調整することによって、容易に各行
25 及び各列で複数のパリティ情報を求めることができる。

次に、本発明の第2の生物学的物質の配列情報の供給方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを保持する供給者(2A)が、そのテキストデータ、又はこれに対応するその数値データを第1ファイル(19)に記録して保

持する第1ステップ（ステップ104）と、その第1ファイルに記録されているそのテキストデータ、又はこのテキストデータに対応するその数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが m ビット（ m は16以上の整数）の部分データに分割し、その

5 部分データの最大値を N_{max} 、この最大値 N_{max} よりも大きい素数を P として、複数行のその部分データに各行毎にその非配列方向にガロア体 $GF(P)$ 上の第1の演算を施して第1組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体 $GF(P)$ 上の第2の演算を施して第2組のパリティ情報を求める第2ステップと、その供給者が、その第1組及び第

10 2組のパリティ情報を第2ファイル（20）に記録して保持する第3ステップと、その生物学的物質の配列情報のユーザ（2B）が、通信回線（1）を介してその供給者よりその第2ファイルに記録されているその2組のパリティ情報を受け取る第4ステップとを有するものである。

この供給方法は、上記の本発明の第2の生物学的物質の配列情報の記録方法を、

15 その配列情報を供給（販売）する際のビジネスモデルに適用したものである。即ち、本発明のビジネスモデルでは、或る生物XのDNAのヌクレオチド、又はタンパク質のアミノ酸などの生物学的物質の配列を最初に解読した供給者は、その配列のテキストデータ（又はこれを変換した数値データ）より、その配列情報を少ないデータ量で近似するパリティ情報を算出し、これをその通信回線を介して

20 ユーザに供給する。本発明においても、ガロア体 $GF(2^m)$ を用いる場合とほぼ同程度にそのパリティ情報は、元のテキストデータに比べて少なくできるため、そのパリティ情報はその通信回線を介して短時間で受信することができる。

本発明の供給方法においても、更にそのユーザが、その2組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内のその供給者の生物学的物質の

25 配列情報との相違部を特定する第5ステップと、この相違部の配列の復元ができない場合に、そのユーザがその通信回線を介してその供給者よりその第1ファイルに記録されているそのテキストデータ、又はその数値データの内のその配列の復元ができない部分の配列情報を受け取る第6ステップとを有することが望ましい。

このようにユーザ側で、パリティ情報だけで検査対象の配列内の供給者の配列との相違部の特定、及び復元ができる場合には、それ以上の配列情報を購入する必要が無い。一方、相違部が多く存在し、パリティ情報のみでは全部の正確なデータが復元できない場合には、例えば復元できない部分のテキストデータ（又は数値データ）のみを購入することによって、通信回線を介して必要な配列情報を短時間に購入できる。従って、通信回線として、携帯電話システムのような比較的低速の通信回線も使用できる。

図面の簡単な説明

- 図1は、本発明の実施の形態の一例で使用されるコンピュータシステムを示す概略構成図である。図2は、その実施の形態の一例で処理対象とするDNA、及びそのヌクレオチドの配列のバイナリーデータによる表現の例を示す図である。図3は、その実施の形態の一例におけるDNA情報の供給者の動作の一部を示すフローチャートである。図4は、図3の動作に続くDNA情報の供給者の動作を示すフローチャートである。図5は、その実施の形態の一例におけるDNA情報のユーザの動作の一部を示すフローチャートである。図6は、図5の動作に続くDNA情報のユーザの動作を示すフローチャートである。図7は、標準試料E（DNA）のヌクレオチド（2048個）の配列を表すテキストデータを4行で32列の部分テキストデータ $T(i, j)$ に分割した状態を示す図である。図8は、標準試料Eの部分データ $A(i, j)$ 、及びこれらから算出されるパリティ $B1(i) \sim C3(j)$ を示す図である。図9は、試料F（DNA）のヌクレオチド（2048個）の配列を表すテキストデータを4行で32列の部分テキストデータ $TF(i, j)$ に分割した状態を示す図である。図10は、試料Fの部分データ $AF(i, j)$ 、及びこれらから算出されるパリティ $B1F(i) \sim C3F(j)$ を示す図である。図11は、標準試料Eのパリティと異なる試料Fのパリティ、及び復元された部分データを示す図である。図12は、未知数 $X1, X2, Y1, Y2$ をガロア体 $GF(2^{128})$ 上で求める場合の計算を示すフローチャートである。図13は、図7の標準試料Eのヌクレオチドの配列を表すテキストデータをバイナリーデータに変換した後、5行で13列の部分データ $B(i,$

j) に分割した状態を示す図である。図 1 4 は、試料 G (タンパク質) のアミノ酸 (8 2 0 個) の配列を表すテキストデータを 4 行で 2 6 列の部分テキストデータに分割した状態を示す図である。図 1 5 は、図 1 4 のアミノ酸の配列に対して計算されたパリティ B 1 G (i) ~ C 3 G (j) を示す図である。

5

発明を実施するための最良の形態

以下、本発明の好ましい実施の形態の一例につき図面を参照して説明する。本例は、所定の DNA (デオキシリボ核酸 : deoxyribonucleic acid) のヌクレオチド (生物学的物質) の配列情報をコンピュータシステムで処理する場合に、本発明を適用したものである。

10

図 1 は、本例のコンピュータシステム 2 A の概略構成を示し、この図 1 において、コンピュータシステム 2 A の中心は、CPU (中央演算処理ユニット)、RAM、ROM 等のメモリ、及びハードディスク装置等の記憶装置等からなる情報処理装置 1 0 である。情報処理装置 1 0 には、ビデオ RAM (VRAM) 1 1 を介して CRT ディスプレイよりなる表示装置 1 2 が接続されると共に、I/O ユニット (入出力装置) 1 4 を介して、記録可能な CD-Recordable ディスク (以下、「CD-R」と言う) 1 6 に対するデータの書き込み、及び CD-R や CD-ROM からのデータの読み込みを行うことができる CD-R/RW ドライブ 1 5 が接続されている。情報処理装置 1 0 には、I/O ユニット 1 4 を介して更に大容量の記憶装置としての記憶容量が数 1 0 0 G バイト程度の磁気ディスク装置 1 7 が接続されている。

15

20

本例の情報処理装置 1 0 中のハードディスク装置には、予め CD-R/RW ドライブ 1 5 を介してオペレーティングシステム、及び後述のように DNA の配列情報を処理するためのアプリケーション・プログラムがインストールされている。また、CD-R 1 6 が本発明の記録媒体に対応しているが、記録媒体としては、CD-R や CD-ROM の他に、フラッシュ ROM、フレキシブルディスク、光磁気ディスク (MO)、デジタルビデオディスク (DVD)、又はハードディスク装置 (例えばインターネットを介して接続できるサーバに備えられたもの) 等を使用することができる。

25

情報処理装置 10 には更に、文字情報の入力装置としてのキーボード 13、ポインティング・デバイス（入力装置）としての光学式のマウス 204、及びルータ（又はモデム等でもよい）よりなる通信制御ユニット 18 が接続されている。マウス 204 は、表示装置 12 の表示画面上のカーソルの位置を指定する信号を発生する変位信号発生部 207、選択すべき情報を指定する信号や各種コマンド等を発生するための左スイッチ 204a 及び右スイッチ 204b（信号発生装置）を備えている。情報処理装置 10、VRAM 11、表示装置 12、キーボード 13、マウス 204、I/O ユニット 14、CD-R/RW ドライブ 15、磁気ディスク装置 17、及び通信制御ユニット 18 等よりコンピュータシステム 2A が構成されている。オペレーティングシステムとして本例では Windows（Microsoft Corporation の登録商標）を使用している。なお、オペレーティングシステムとして、それ以外の UNIX（X/Open の登録商標）、OS/2（IBM Corporatin の登録商標）、Mac OS（Apple Computer の登録商標）、又は Linux（Linus Torvalds の商標又は登録商標）等を使用する場合にも本発明が適用できることは言うまでもない。

そして、コンピュータシステム 2A（情報処理装置 10）は、通信制御ユニット 18 を介して一般電話回線よりなる通信ネットワーク 1 に接続され、通信ネットワーク 1 には各種コンテンツのプロバイダ 3、及び別のコンピュータシステム 2B、及び不図示の多くのサーバやコンピュータシステムが接続されている。また、本例のコンピュータシステム 2A、2B 及びプロバイダ 3 は、通信ネットワーク 1 を介するインターネットによって相互に接続されている。この場合、コンピュータシステム 2A の所有者が DNA 情報の供給者（販売者）であり、コンピュータシステム 2B の所有者がその DNA 情報のユーザ（購入者）である。そして、後者のコンピュータシステム 2B には、予め前者のコンピュータシステム 2A と同様の DNA の配列情報を処理するためのアプリケーション・プログラムがインストールされている。

さて、本例のコンピュータシステム 2A の情報処理装置 10 には、I/O ユニット 14 を介して、生物学的物質としての DNA 中の一系列のヌクレオチドの配列（又は塩基の配列）を読み取るための配列読み取り装置としてのシーケンサー

(DNA Sequencer) 4が接続されている。シーケンサー 4は、一例としてサンガーの方法 (Sanger method) によってDNAを構成する1対の重合体の鎖の一方の鎖のヌクレオチドの配列を読み取る。サンガーの方法は、例えば文献1 (Maxim D. Frank-Kamenetskii: Unraveling DNA (the most important molecule of life, revised and updated), translated by Lev Liapin, Chapter 6 (pp. 59-70) (Pers
5 eus Books, 1997)) に開示されている。シーケンサー 4は、読み取った一列のヌクレオチドの配列をテキストデータ形式で内部の大容量の記憶装置に記憶すると共に、情報処理装置 10からの要求に応じて、その記憶装置中の所定のヌクレオチドの配列のテキストデータをI/Oユニット 14を介して情報処理装置 10に
10 供給する。これに対して情報処理装置 10は、DNAの配列情報を処理するためのアプリケーション・プログラムに基づいて以下の処理を行う。なお、シーケンサー 4の代わりに、DNA及びRNA (リボ核酸: ribonucleic acid) 等の核酸を構成する一列のヌクレオチドの配列 (又は塩基の配列) の情報のデータベースを接続してもよい。

15 先ず、本例の情報処理装置 10の第1の基本的な処理動作につき説明する。情報処理装置 10は、シーケンサー 4から供給される所定のDNAのヌクレオチドの配列を示すテキストデータ (本例ではアスキーコード (ANSI形式) を用いる) を磁気ディスク装置 17中のマスターファイル 19にそのまま記録すると共に、そのテキストデータをよりデータ量の少ない数値データに変換し、この変換
20 後の数値データを磁気ディスク装置 17中のワーキングファイル 20に記録する。なお、以下の説明において、2進数表示の数kは bin(k) で、16進数表示の数kは hex(k) で表すものとする。

この場合、DNAは4種類のヌクレオチドより構成されており、シーケンサー 4から供給されるテキストデータ中では、塩基としてアデニン (adenine) を含む
25 ヌクレオチド、グアニン (guanine) を含むヌクレオチド、シトシン (cytosine) を含むヌクレオチド、及びチミン (thymine) を含むヌクレオチドがそれぞれ文字A, G, C, 及びTで表されている。そして、文字A, G, C, 及びTには、データ上ではそれぞれhex(41), hex(47), hex(43), hex(54) よりなる1バイト (8ビット) のアスキーコードが割り当てられている。また、RNAの場合には、チミン

を含むヌクレオチドの代わりにウラシル (uracil) を含むヌクレオチドが、文字 U (hex (55)) で表されている。従って、n 個のヌクレオチドの配列を示すテキストデータのデータ量は n バイトとなる。なお、それらの n 個のヌクレオチドの配列は、n 個の塩基 (アデニン、グアニン、シトシン、チミン (又はウラシル)) の配列ともみなすことができる。

本例ではそのテキストデータを、情報量を少なくすることなく最も少ないデータ量で表すために、DNA 中の 4 種類のヌクレオチドを互いに異なる 2 ビットのデータで表す。この際に、DNA においては、1 対の塩基 (アデニン及びチミン) が互いに相補的であり、別の 1 対の塩基 (グアニン及びシトシン) が互いに相補的である。そこで、相補的な塩基を含む 1 対のヌクレオチドを互いに相補的であるとして、1 対の互いに相補的なヌクレオチド、即ちアデニンを含むヌクレオチド及びチミンを含むヌクレオチドに、互いにビット反転の関係にある 1 対のデータを割り当て、別の 1 対の互いに相補的なヌクレオチド、即ちグアニンを含むヌクレオチド及びシトシンを含むヌクレオチドに、互いにビット反転の関係にある別の 1 対のデータを割り当てる。本例ではそのデータの割り当てとして表 1 (変換テーブル) を用いる。なお、表 1 は、ヌクレオチドの配列を示すテキストデータ中の文字 A, T (又は U), G, C, をそれぞれ bin (00), bin (11), bin (01), bin (10) で置換することを意味している。

《表 1》

ヌクレオチド	2 ビットのデータ
アデニンを含むヌクレオチド (A)	bin (00)
チミン (ウラシル) を含むヌクレオチド (T 又は U)	bin (11)
グアニンを含むヌクレオチド (G)	bin (01)
シトシンを含むヌクレオチド (C)	bin (10) 。

なお、本例では各ヌクレオチドを 2 ビットのデータで表しているが、これは各塩基を 2 ビットのデータで表すのと等価である。また、データの割り当ては表 1 には限定されず、例えばチミンを含むヌクレオチドを bin (00) 、アデニンを含むヌクレオチドを bin (11) とするか、又はグアニンを含むヌクレオチドを bin (10) 、シトシンを含むヌクレオチドを bin (01) としてもよい。それ以外に、アデニンを

含むヌクレオチド及びチミンを含むヌクレオチドに、1対のデータbin(01), bin(10)を割り当て、グアニンを含むヌクレオチド及びシトシンを含むヌクレオチドに1対のデータbin(00), bin(11)を割り当てるようにしてもよい。また、RNAの場合には、チミンを含むヌクレオチドに割り当てられているデータをウラシルを含むヌクレオチドに割り当てて、それ以外のヌクレオチドにはDNAのヌクレオチドと同じデータを割り当てればよい。

本例では図2に示すDNA分子5のヌクレオチドの配列情報を扱うものとする。その配列情報は、NCBI(The National Center for Biotechnology Information)より提供されているウェブサイト1(<ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/>)より入手した大腸菌(*Escherichia coli*: *E. coli*)のDNAの1列のヌクレオチドの配列の一部である。

図2において、DNA分子5は、1対の重合体の鎖6A、6B(二重らせん)より構成され、一方の重合体の鎖6Aは、アデニンを含むヌクレオチド7A、グアニンを含むヌクレオチド7G、シトシンを含むヌクレオチド7C、及びチミンを含むヌクレオチド7Tよりなる4種類のヌクレオチドの配列であり、他方の重合体の鎖6Bは、鎖6Aに対して相補的なヌクレオチドの配列である。この際に、図1の情報処理装置10には一方の重合体の鎖6Aの配列を示すテキストデータ、即ち”AGCTTT・・・”の文字列のデータが供給される。それに対して、情報処理装置10は、そのテキストデータ中の文字A、G、C、Tを表1の変換テーブルに基づいて順次2ビットのデータに変換することによって、数値データとしてのバイナリーデータBNA(=bin(0001101111・・・))を得る。そして、このバイナリーデータBNAが図1の磁気ディスク装置17のワーキングファイル20に記録される。そのバイナリーデータBNAのデータ量は、元のテキストデータの1/4となっている。

25 この場合、そのワーキングファイル20の先頭の所定数のバイトの領域に、例えばその配列がDNA又はRNAのどちらかを示すデータ(即ち、bin(11)を文字T又は文字Uの何れに解釈するかを示すデータ)、ヌクレオチドの個数を示すデータ、及びその他の必要なデータを記録しておけばよい。また、そのワーキングファイル20の長さが1バイト(8ビット)単位で規定されている場合に、バ

イナリーデータBNAの末尾で1バイトの端数のデータが生じたときには、予め定めておいたダミーデータを付加すればよい。これでもデータ量は殆ど増加しない。そして、一例としてユーザ（コンピュータシステム2Bの所有者）から供給者（コンピュータシステム2Aの所有者）に対して図2のDNA分子5の配列情報5の購入希望が届いたときに、ワーキングファイル20のデータが通信ネットワーク1及び不図示のプロバイダを介して、電子メールの添付ファイルとしてコンピュータシステム2B側に供給される。この際に、そのワーキングファイル20のデータを更に圧縮ファイル（ZIPファイル、又はLHAファイル等）として送信してもよい。この際に、ワーキングファイル20のデータ量はもとのテキストデータ10のほぼ1／4であるため、元のテキストデータ（更に圧縮ファイルとした場合も同様）自体を送信する場合に比べて送信時間はほぼ1／4となり、供給者側及びユーザ側双方の通信コストが低減できる。

そして、ユーザ側で、その受信したワーキングファイル20のデータから図2の一方の重合体の鎖6Aの配列のテキストデータを復元する場合には、コンピュータシステム2Bにおいて、ワーキングファイル20中のバイナリーデータBNAを、表1を用いて文字A、G、C、T（又はU）の何れかに順次逆変換すればよい。また、その際に例えば図2の他方の相補的な重合体の鎖6Bのヌクレオチドの配列を示すテキストデータが必要になった場合には、コンピュータシステム2Bにおいて、図2に示すように、バイナリーデータBNAのビット毎の反転操作を行って反転バイナリーデータNOT(BNA)（=bin(1110010000···)）を得る。この反転バイナリーデータNOT(BNA)は、他方の重合体の鎖6Bのヌクレオチドの配列を示すテキストデータ（文字列”TCGAAA···”）を表1に従って変換したバイナリーデータBNBと同一である。従って、その反転バイナリーデータNOT(BNA)を、表1を用いて文字A、G、C、T（又はU）の何れかに順次逆変換するのみで、極めて高速に相補的な重合体の鎖6Bの配列のテキストデータを25得ることができる。この際に、通常のコンピュータにおいては、ビット毎の反転操作は、極めて高速に実行することができる。なお、そのビット毎の反転操作は、例えばbin(111111···)との排他的論理和演算で代用してもよい。

なお、ワーキングファイル20のデータを通信ネットワーク1を介してユーザ

側に送信する代わりに、ワーキングファイル20の内容をCD-R/RWドライブ15によってCD-R16に記録し、このCD-R16を郵送等によってユーザ側に供給してもよい。例えば一人の人間の全部のDNAの配列情報（ゲノム）は、テキストデータでは3Gバイト程度になるが、これを表1を用いて本例の数値データとしてのバイナリーデータに変換すると、3/4Gバイト程度、即ち750Mバイト程度になる。現在のCD-R、CD-ROMの容量は約650Mバイトであるため、その750Mバイト程度のバイナリーデータは例えば一部又は全部を圧縮ファイルとすることによって、余裕を持ってCD-R16に記録することができる。これに対して、その750Mバイト程度のデータを通信ネットワーク1を介して送信しようとする、現状でも送信時間がかかり過ぎる場合がある。

また、一つのアミノ酸の種類は一系列の3個のヌクレオチドの配列、即ち一つの遺伝子コドン（codon）によって決定される。そこで、一つのアミノ酸に対応する3個のヌクレオチドをそれぞれ2ビットのデータで表したときに得られる6ビットのデータの内で、最も小さいデータでそのアミノ酸を表すようにしてもよい。この際に、個々のデータは、1バイト単位が扱い易いため、その6ビットのデータの前後に2ビットの識別データを付加して得られる1バイトのデータで一つのアミノ酸を表すようにしてもよい。これによって、ヌクレオチドとアミノ酸とで共通のコードを使用できる利点がある。

次に、本例の情報処理装置10の第2の基本的な処理動作につき説明する。本例では、ヌクレオチドの配列を示す膨大な量のテキストデータ（又はこれを表1に基づいて変換して得られる数値データ）より、所定のハッシュ関数を用いて数学的な要約値（message digest）を算出する。本例ではそのハッシュ関数として、ライベスト（R. Rivest）によって提案されたMD5ハッシュ関数を使用する。MD5ハッシュ関数のアルゴリズムについては、ネットワークワーキンググループ及びライベストによって開設されているウェブサイト2 (<http://www.kleinsmidt.com/edi/md5.htm>) に開示されている。また、そのMD5ハッシュ関数のアルゴリズムは、国際公開公報W01/80431 A1 にも開示されている。また、2002年10月16日付け提出の米国出願第10/272, 107号で開示されている

MD 5 ハッシュ関数及びその他のハッシュ関数のアルゴリズムは、そっくりそのまま引用して本願に組み込まれている。或るテキストデータ（テキストファイル）にそのMD 5 ハッシュ関数を施すことによって、128 ビットの要約値が得られる。通常のコンピュータでも今後は処理単位が64 ビットのCPUが使用されるようになると考えられるが、この場合に128 ($= 2 \cdot 64$) ビットの要約値は非常に扱い易い長さである。この場合には、192 ($= 3 \cdot 64$) ビットの要約値も比較的扱い易いと考えられる。

また、本例では、そのMD 5 ハッシュ関数のプログラムとして、そのウェブサイト2において公開されている、RSAデータセキュリティー社（RSA Data Security Inc.）によって開発されたプログラムを使用した。

その要約値の使用方法の一例として、DNAの配列情報の供給者（情報処理装置10）は、所定の生物のDNAのヌクレオチドの配列を読み取り、これに対応するテキストデータより、上記のハッシュ関数を用いて要約値を算出し、この要約値をその生物の名称、及びDNAの位置を示す情報と共にインターネット上で閲覧可能にする。これによって、その供給者は、そのテキストデータ自体を公開することなく、その生物のDNAの配列情報を最先に解読したことを主張できると考えられる。その後、或るユーザからのその配列情報の購入希望が来たときに、その供給者は、そのヌクレオチドの配列のテキストデータを表1を用いてバイナリーデータに変換し、このバイナリーデータを例えば通信ネットワーク1を介して電子メールの形でそのユーザに送信する。これに対してユーザ側では、そのバイナリーデータを表1を用いてテキストデータに変換し、この変換されたテキストデータに上記のハッシュ関数を施して要約値を求める。

そして、この要約値とその供給者によって公開されている要約値とが等しいときには、購入した配列情報が、供給の保持している配列情報と等しいことが極めて高い確率で保証される。更に、ユーザ側では、複数の供給者が公開している要約値を比較することによって、同じ配列情報を異なる複数の供給者から重複して購入することを防止することができる。これらの際に、ヌクレオチドの配列の長さ、及び先端部や末尾の一部の短い配列の比較を行うことによって、その配列情報の同一性を高めることができる。

なお、ハッシュ関数としては、例えば文献 2 (FIPS Publication 180, 1993) で開示されているように、NBS (National Bureau of Standards) によって提案された SHS (Secure Hash Standard) ハッシュ関数を使用してもよい。SHS ハッシュ関数は、MD5 ハッシュ関数よりも複雑な演算を行うと共に、160 ビットの要約値が得られる。これに関して、例えばタンパク質を構成するアミノ酸の配列数は20個～1000個程度であり、特に一文字表記を使用する際にはそれに対応するテキストデータも20バイト～1kバイト程度に短くなるため、要約値から元のテキストデータが推定し易いと考えられる。そこで、アミノ酸の配列情報の要約値を求める際には、SHS ハッシュ関数を使用する方が望ましいことがある。

また、例えばヌクレオチドの配列を示す2つの膨大な長さのテキストデータの同一性を確認するために、ハッシュ関数の要約値を算出するような場合には、それ程複雑な計算を繰り返して行う必要は無いと考えられる。そこで、このような用途では、例えば文献 3 (R. L. Rivest: "The MD4 message digest algorithm", Lecture Notes in Computer Science, 537, 303-311 (1991)) で開示されている MD4 ハッシュ関数を使用してもよいと考えられる。また、そのように単に同一性を確認する用途では、要約値の長さも40ビット～128ビット程度でよい場合がある。

次に、本例のDNA情報の供給者（コンピュータシステム2A）と、ユーザ（コンピュータシステム2B）との間でDNAの配列情報を受け渡す際のビジネスモデルの一例につき図3～図6のフローチャートを参照して詳細に説明する。まず、DNA情報の供給者側では、図3のステップ101において、シーケンサー4を使用して標準となる試料（標準試料Eとする）のDNA中の一方の系列のヌクレオチドの配列を読み取り、読み取った配列を表すテキストデータTX1を情報処理装置10に供給する。本例では、その標準試料Eを大腸菌として、そのテキストデータTX1として、図7に示すように、上記のウェブサイト1から入手した大腸菌のDNAの配列情報の内の、最初から2048個までのヌクレオチドの配列を示すテキストデータを使用する。

標準試料EのDNA配列は配列番号1に示されている。図7のテキストデータ

は、配列番号 1 の配列から数字データを除いて、a, g, c, t の文字をそれぞれ A, G, C, T で置き換えたものに相当する。

次のステップ 102 において、情報処理装置 10 は、供給されたテキストデータ TX1 に上記の MD5 ハッシュ関数を施して 128 ビットの要約値 AB1 を求めると共に、そのヌクレオチドの配列の数 NA1、及び先頭と末尾との 8 個ずつのヌクレオチドの配列 ST1, SB1 を求める。テキストデータ TX1 に対する具体的な値は下記の通りである。

AB1 = hex (849339ac244cde42b5346ab5989aab61) ... (11)

NA1 = 2048

10 ST1 = AGCTTTTC, SB1 = CGCGAAGG

次のステップ 103 において、情報処理装置 10 は、テキストデータ TX1 を逆方向に並べ替えたテキストデータ TXR1 (=GGAAGC...TTTCGA) を求め、このテキストデータ TXR1 の MD5 ハッシュ関数による要約値 ABR1、及びこのテキストデータ TXR1 の先頭と末尾との 8 個ずつのヌクレオチドの配列 STR1, SBR1 を求める。配列 STR1, SBR1 は、上記の配列 SB1, ST1 をそれぞれ逆方向に並べ替えることによって容易に求めることができる。これらの具体的な値は以下の通りである。

ABR1 = hex (4eb1feae30f522642b912ce3ea09652b) ... (12)

STR1 = GGAAGCGC, SBR1 = CTTTTTCGA

20 次のステップ 104 において、情報処理装置 10 は、標準試料 E の名前の情報 (試料を特定する情報)、配列の数 NA1、テキストデータ TX1、配列 ST1, SB1、要約値 AB1、逆方向の配列 STR1, SBR1、及び逆方向の要約値 ABR1 を磁気ディスク装置 17 のマスターファイル 19 に記録する。この際に、マスターファイル 19 を複数のファイルとして、テキストデータ TX1 と、それ
25 以外のデータとを別のファイルに記録してもよい。また、テキストデータ TX1 が例えば 100M バイト程度以上になる場合には、テキストデータ TX1 を複数のマスターファイルに分割して記録してもよい。

次のステップ 105 において、情報処理装置 10 は、図 7 に示すように、標準試料 E のテキストデータ TX1 を配列方向 (ヌクレオチドの配列方向) に N 行で、

その配列方向に直交する方向（以下、「非配列方向」という）にM列の16文字の長さの部分テキストデータ $T(i, j)$ ($i = 1 \sim N$, $j = 1 \sim M$)に分割する。なお、N、Mはそれぞれ2以上の任意の整数であり、(4)式、(5)式を用いて既に説明したように、テキストデータTX1が100kバイト程度（又はこの整数倍）であるときに、このテキストデータTX1に対して1/20程度のデータ量のパリティ情報を得たい場合には、例えばNの値が64、Mの値が128に設定される。以下では説明を簡単にするために、図7に示すようにテキストデータTX1を4行で、かつ32列に分割した場合を想定する。即ち、 $N = 4$, $M = 32$ とする。この場合、本例では端数は生じないが、例えば図7において、最後の部分テキストデータ $T(4, 32)$ 中の文字が16個より少ない場合には、足りない部分には予め定めた文字（例えば文字A）をダミーデータとして付加すればよい。また、部分テキストデータ $T(i, j)$ の長さは、16文字以外の任意の長さでよいが、処理速度を高めるためには、8文字の倍数が効率的である。

更に、情報処理装置10は、図7の16文字分の各部分テキストデータ $T(i, j)$ をそれぞれ所定の変換テーブルに従って128 ($= 16 \times 8$)ビットのバイナリーデータ（数値データ）よりなる部分データ $A(i, j)$ に変換する。本例ではその変換テーブルとして、次のように部分テキストデータ $T(i, j)$ を単にアスキーコードに変換する関数 $asc(T(i, j))$ を用いる。

$$A(i, j) = asc(T(i, j)) \quad \dots (13)$$

なお、図7に $T(3, 11)$ の変換例で示すように、関数 $asc(T(i, j))$ は、部分テキストデータ $T(i, j)$ の先頭の文字のコードが最下位桁となり、末尾の文字のコードが最上位桁となるように変換を行う。この際に、例えば最後の列の部分データ $A(i, 32)$ が128ビットにならないときには、その上位に予め定めた文字コード、又は数値データの0 (hex(000...))などのダミーデータが付加される。この結果、図8に示す4行で、32列の部分データ $A(i, j)$ が得られる。また、部分データ $A(i, j)$ を対応するヌクレオチドの配列方向に連続して配列したときの集合体（数値データ）をバイナリーデータBN1とする。図7の部分テキストデータ $T(i, j)$ と図8の部分データ $A(i, j)$ とは実質的に同じデータ量である。

次に、本例では、部分データ $A(i, j)$ をガロア体 (Galois field) $GF(2^m)$ 上の元のベクトル表示とみなして、部分データ $A(i, j)$ に対してガロア体 $GF(2^m)$ 上の所定の演算を施す。本例の部分データ $A(i, j)$ は 128 ビットであるため、 m の値は 128 (64 の 2 倍) となり、ガロア体 $GF(2^{128})$ が使用される。また、本例ではガロア体 $GF(2^{128})$ 上の既約多項式 $GF(X)$ 及び生成元 α として次の式を使用する。なお、ガロア体 $GF(2^m)$ は、拡大ガロア体とも呼ばれることがある。

$$GF(X) = 1 + X^{121} + X^{126} + X^{127} + X^{128} \quad \cdots (14)$$

$$\alpha = X \quad \cdots (15)$$

ガロア体 $GF(2^{128})$ 上のベクトル表示では、 $GF(X)$ は $\text{bin}(1110000100 \cdots 01)$ となり、 α は $\text{bin}(00 \cdots 0010)$ となる。なお、生成元 α としては、 $(1 + X)$ など也可以使用できる。また、ガロア体 $GF(2^{128})$ 上の既約多項式としては、例えば次の既約多項式 $GF'(X)$ も使用でき、この既約多項式 $GF'(X)$ に対する生成元としては次の α' などを使用できる。

$$GF'(X) = 1 + X^{11} + X^{124} + X^{125} + X^{126} + X^{127} + X^{128} \quad \cdots (14A)$$

$$\alpha' = 1 + X + X^2 \quad \cdots (15A)$$

また、多項式 $GF(X)$ が既約かどうかを確かめる最も簡単な方法は、その多項式 $GF(X)$ の次数を m として、 $m/2$ を超えない整数を m' とすると、その多項式 $GF(X)$ を次数 m' 以下の全ての多項式で割ってみることである。このときに、割り切れる多項式がなければ、多項式 $GF(X)$ は既約である。

また、特に次数 m の大きい多項式 $GF(X)$ が既約であることは、例えば文献 4 (Van der Waerden, B. L. (1953), Modern Algebra (2 vols.), p. 77, Ungar, New York) に記載してある「Kroneckerの方法」で確認することができる。また、 $GF(X)$ が既約であることは、実用的には、ウェブサイト 3 (<http://archives.math.utk.edu/software/msdos/number.theory/ubasic/.html>)、又はウェブサイト 4 (<http://www.rkmath.rikkyo.ac.jp/~kida/ubasic.htm>) に開示されている整数論研究用のソフトウェアである「UBASIC」中の組み込み関数「POLFACT2」を用いても確認することができる。

また、ガロア体 $GF(2^m)$ 上の生成元 α は、 $k = 2^m - 1$ とおくと、既約

多項式GF (X) を法として、次の関係を満たす。

$$\alpha^k = 1 \pmod{GF(X)} \quad \cdots (16)$$

$$\alpha^{k'} \neq 1 \pmod{GF(X)} \quad (1 \leq k' < k) \quad \cdots (17)$$

そこで、素数 p_1, p_2, \dots, p_r 及び整数 n_1, n_2, \dots, n_r を用いて、

5 k が次のように因数分解できるものとする。

$$k = 2^m - 1 = p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r} \quad \cdots (18)$$

このとき、生成元 α とは、既約多項式GF (X) を法として、 α の $(p_1^{n_1-1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r})$ 乗、 $(p_1^{n_1} \cdot p_2^{n_2-1} \cdot \dots \cdot p_r^{n_r})$ 乗、 \dots 、 $(p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r-1})$ 乗が何れも1とならないものであればよい。

10 また、ガロア体GF (2^m) 上の任意の0以外の元 β についても(16)式が成立するため、 $k (= 2^m - 1)$ を用いて β の逆元 β^{-1} は次のように計算することも可能である。

$$\beta^{-1} = \beta^{k-1} \pmod{GF(X)} \quad \cdots (16R)$$

従って、例えば部分データA (i, j) を β で除算する場合には、部分データ

15 A (i, j) に β^{k-1} を乗算すればよい

次のステップ106において、情報処理装置10は、図8の各行 ($i = 1 \sim 4$) の部分データA (i, j) に対してガロア体GF (2^{128}) 上で、非配列方向 ($j = 1 \sim 32$) に対する和である第1パリティ (Parity) B1 (i)、 $\sum \alpha^{(j-1)} \cdot A (i, j)$ である第2パリティB2 (i)、及び $\sum \alpha^{2(j-1)} \cdot A (i, j)$ である第3パリティB3 (i) を計算する。これらの非配列方向のパリティB1 (i) \sim B3 (i) (第1組のパリティ情報) は、生成元 α を用いて、かつ既約多項式GF (X) を法として以下のように表すことができる。パリティB1 (i) \sim B3 (i) における記号 (Σ) は係数 j に対する $1 \sim 32$ の和を意味しており、以下の式は係数 i の $1 \sim 4$ の範囲で計算される。

$$\begin{aligned} 25 \quad B1 (i) &= \Sigma A (i, j) = A (i, 1) \\ &\quad + A (i, 2) + \dots + A (i, 32) \quad \cdots (19) \end{aligned}$$

$$\begin{aligned} B2 (i) &= \Sigma \alpha^{(j-1)} \cdot A (i, j) = A (i, 1) \\ &\quad + \alpha \cdot A (i, 2) + \dots + \alpha^{31} \cdot A (i, 32) \quad \cdots (20) \end{aligned}$$

$$B3 (i) = \Sigma \alpha^{2(j-1)} \cdot A (i, j) = A (i, 1)$$

$$+ \alpha^2 \cdot A(i, 2) + \dots + \alpha^{62} \cdot A(i, 32) \quad \dots (21)$$

この場合、(19)式のパリティ $B1(i)$ のベクトル表示は、部分データ $A(i, j)$ についてビット毎に排他的論理和演算を行って得られる結果と同じである。また、(20)式、(21)式のパリティ $B2(i)$ 、 $B3(i)$ は、それぞれ部分データ $A(i, j)$ を(1)式のように127次 ($m=128$) 以下の多項式で表して、既約多項式 $GF(X)$ を法として演算を行うことによって計算することができる。

更に、情報処理装置10は、図8の各列 ($j=1 \sim 32$) の部分データ $A(i, j)$ に対してガロア体 $GF(2^{128})$ 上で、配列方向 ($i=1 \sim 4$) に対する和である第1パリティ $C1(j)$ 、 $\sum \alpha^{(i-1)} \cdot A(i, j)$ である第2パリティ $C2(j)$ 、及び $\sum \alpha^{2(i-1)} \cdot A(i, j)$ である第3パリティ $C3(j)$ を計算する。これらの配列方向のパリティ $C1(j) \sim C3(j)$ (第2組のパリティ情報) は、生成元 α を用いて、かつ既約多項式 $GF(X)$ を法として以下のように表すことができる。パリティ $C1(j) \sim C3(j)$ における記号 (\sum) は係数 i に対する $1 \sim 4$ の和を意味しており、以下の式は係数 j の $1 \sim 32$ の範囲で計算される。

$$C1(j) = \sum A(i, j) = A(1, j) + A(2, j) + \dots + A(4, j) \quad \dots (22)$$

$$C2(j) = \sum \alpha^{(i-1)} \cdot A(i, j) = A(1, j) + \alpha \cdot A(2, j) + \dots + \alpha^3 \cdot A(4, j) \quad \dots (23)$$

$$C3(j) = \sum \alpha^{2(i-1)} \cdot A(i, j) = A(1, j) + \alpha^2 \cdot A(2, j) + \dots + \alpha^6 \cdot A(4, j) \quad \dots (24)$$

部分データ $A(i, j)$ に対して実際にパリティ $B1(i) \sim B3(i)$ 、及びパリティ $C1(j) \sim C3(j)$ を計算した結果のベクトル表示が、図8に16進数表示で示されている。この例においては、各行のパリティ $B1(i) \sim B3(i)$ 、及び各列のパリティ $C1(j) \sim C3(j)$ はそれぞれ3個であるため、2つのヌクレオチドの配列の比較を行う場合に、各行及び各列において、それぞれ3個までの部分データ $A(i, j)$ の相違部の復元を正確に行うことができる。従って、各行及び各列において、部分データ $A(i, j)$ の相違部の位置

の検出（特定）だけを行うと共に、相違部の復元を1個だけ行えばよい場合には、パリティ情報として、 $B_1(i)$ 及び $C_1(j)$ 、又は $B_2(i)$ 及び C_2

(j) だけを使用（計算）するだけでもよい。後者のパリティ $B_2(i)$ 、 $C_2(j)$ だけを使用する場合には、例えば或る行又は列で2つの部分データ $A(i, j)$ が入れ替わったような配列であっても、配列の相違部の位置を検出できる利点がある。

また、各行及び各列において、それぞれ2個までの部分データ $A(i, j)$ の相違部の復元を正確に行うことができればよい場合には、第1組のパリティ情報として $B_1(i)$ 、 $B_2(i)$ 、 $B_3(i)$ の何れか2つ、及び第2組のパリティ情報として $C_1(j)$ 、 $C_2(j)$ 、 $C_3(j)$ の何れか2つだけを使用（計算）するだけでもよい。また、各行と各列とで復元できる部分データの個数が違ってもよい場合には、第1組のパリティ情報と第2組のパリティ情報とでパリティの個数が違ってもよい。更に、各行又は各列において、それぞれ4個以上の相違部の復元を正確に行うためには、例えば $\sum \alpha^{s(j-1)} \cdot A(i, j)$ であるパリティ $B_s(i)$ ($s=4, 5, \dots$)、又は $\sum \alpha^{t(i-1)} \cdot A(i, j)$ であるパリティ $C_t(j)$ ($t=4, 5, \dots$) を計算すればよい。

また、図8の部分データ $A(i, j)$ の配列が実際には、64行×128列であるとする、図8の例のように、各行及び各列で3個までの相違部の復元を行う場合には、パリティ $B_1(i) \sim B_3(i)$ 、 $C_1(j) \sim C_3(j)$ はそれぞれ128ビット（16バイト）であるため、全部のパリティ情報のデータ量は、 $576 \cdot 16 (= (64 + 128) \cdot 3 \cdot 16)$ バイトとなる。一方、部分データ $A(i, j)$ のデータ量は、 $8192 \cdot 16 (= 64 \cdot 128 \cdot 16)$ バイトとなる。従って、全部のパリティ情報のデータ量は、全部の部分データ (i, j) に対してほぼ1/14程度に減少している。

次に図4のステップ107において、情報処理装置10は、標準試料Eの名前の情報、配列の数 N_A 、バイナリーデータ B_{N1} 、パリティ $B_1(i) \sim B_3(i)$ 、 $C_1(j) \sim C_3(j)$ を磁気ディスク装置17のワーキングファイル20に記録する。この際に、ワーキングファイル20を複数のファイルとして、バイナリーデータ B_{N1} と、パリティ $B_1(i) \sim B_3(i)$ 、 $C_1(j) \sim C$

3 (j) とを別のファイルに記録してもよい。更に、バイナリーデータBN1と共に、ステップ102で算出した要約値AB1をワーキングファイル20に記録してもよい。

また、バイナリーデータBN1が長いときには、バイナリーデータBN1を複数のファイルに分割して記録してもよい。更に、図7のテキストデータTX1 (ひいては図8のバイナリーデータBN1) がかなり長い場合には、テキストデータTX1を例えば数100kバイト程度を単位として複数のデータ群に分割し、各データ群毎にパリティB1(i)~B3(i), C1(j)~C3(j)を求めるようにしてもよい。

更に、ステップ107において、DNA情報の供給者は、ワーキングファイル20に記録した情報、即ち標準試料Eの名前の情報、配列の数NA1、バイナリーデータBN1、パリティB1(i)~B3(i), C1(j)~C3(j)と、マスターファイル19に記録した要約値AB1, ABR1の情報とを、CD-R/RWドライブ15を介してCD-R16に記録してもよい。このCD-R16から、更に多数のCD-ROMを作製してもよく、これらの記録媒体が郵送等によってユーザに販売される。

次の、ステップ108において、情報処理装置10は、標準試料Eの名前の情報、配列の数NA1、配列ST1, SB1、要約値AB1、逆方向の配列STR1, SBR1、及び逆方向の要約値ABR1を磁気ディスク装置17のコンテンツファイル21に記録する。図7のテキストデータTX1が仮に100Mバイト程度の膨大なものであっても、コンテンツファイル21に記録されるデータは500バイト程度の僅かなものである。更に、情報処理装置10は、コンテンツファイル21中の情報を通信ネットワーク1を介してコンテンツのプロバイダ3に送信する。これによって、コンテンツファイル21中の情報はプロバイダ3のサーバ内の閲覧可能なコンテンツファイル31に記録されて、第3者がインターネットを介して自由に閲覧できるようになる。

次のステップ109において、DNA情報の供給者は、ユーザから購入要求が来るのを待つ状態となる。そして、(a) ユーザから標準試料Eに対する簡易データの要求があったときには、ステップ110に移行して、情報処理装置10は、

磁気ディスク装置 17 のワーキングファイル 20 中のパリティ情報（パリティ B 1 (i) ~ B 3 (i) , C 1 (j) ~ C 3 (j) ）を例えば電子メールの添付ファイルとしてそのユーザに送信する。一方、ステップ 109 において、(b) ユーザから完全データの要求があったときには、ステップ 111 に移行して、情報処理装置 10 は、ワーキングファイル 20 中のバイナリーデータ B N 1 を Z I P ファイル等の形式で圧縮し、この圧縮されたデータを例えば電子メールの添付ファイルとしてそのユーザに送信する。この際に必要に応じて、ハッシュ関数による要約値 A B 1 を同時に送信してもよい。本例によれば、簡易データ（パリティ情報）はデータ量が少ないために短時間で送信することができる。

10 また、ステップ 109 において、ユーザは、必要に応じて部分データ、即ち図 8 の全部の部分データ A (i , j) の内の所望のデータ、例えば 2 つの部分データ A (4 , 16) 及び A (1 , 17) のみをその供給者から購入するようにしてもよい。これによって、必要な正確なデータのみを短時間に入手することができる。

15 次に、DNA 情報のユーザ（図 1 のコンピュータシステム 2 B の所有者とする）側では、図 5 のステップ 121 において、図 1 の通信ネットワーク 1（インターネット）を介してプロバイダ 3 のサーバ内のコンテンツファイル 31 の内容を閲覧し、その中からステップ 108 で供給者（図 1 のコンピュータシステム 2 A）から送信された情報、即ち標準試料 E の名前の情報、ヌクレオチドの配列の数 N A 1、配列 S T 1, S B 1、要約値 A B 1、逆方向の配列 S T R 1, S B R 1、及び逆方向の要約値 A B R 1 を読み取り、読み取った情報をコンピュータシステム 2 B 内の記憶装置の一時ファイルに記録する。

25 次の、ステップ 122 において、そのユーザは、不図示の DNA のシーケンサーを用いて、標準試料 E と同じ種類で検査対象の試料 F の DNA 中の一方の系列のヌクレオチドの配列を読み取り、読み取られた配列を示すテキストデータ T X 2（アスキーコードとする）をコンピュータシステム 2 B 内の情報処理装置に取り込む。その検査対象の試料 F とは、例えば突然変異を起こしていると思われる大腸菌であり、そのテキストデータ T X 2 は、標準試料 E のテキストデータ T X 1 と同様に最初から 2048 個までのヌクレオチドの配列を示すものとする。

試料 F の DNA 配列は配列番号 2 に示されている。後述の図 9 のテキストデータは、配列番号 2 の配列から数字データを除いて、a, g, c, t の文字をそれぞれ A, G, C, T で置き換えたものに相当する。

図 9 は、その試料 F の DNA のヌクレオチドの配列に対応するテキストデータ TX 2 を示し、この図 9 の配列の内のアンダーラインを付した部分のみが、図 7 の標準試料 E の配列と異なっている。即ち、試料 F の配列は、標準試料 E の部分テキストデータ T (4, 16), T (1, 17) の部分だけが以下のように異なっている。なお、この段階では、ユーザは、試料 F の配列と標準試料 E の配列とのどの部分が相違しているのかは分からない。

10	標準試料 E	試料 F
	T (4, 16) = ATTTGGACGGACGTTG → ATTTGGAC <u>ATTATGGC</u>	
	T (1, 17) = ACGGGGTCTATACCTG → <u>GGCCAACTT</u> ATACCTG	

そして、ユーザのコンピュータシステム 2 B 側の情報処理装置においても、DNA の配列情報を処理するためのアプリケーション・プログラムが起動されている。そして、その情報処理装置は、ステップ 1 2 3 において、読み取られたテキストデータ TX 2 に上記の MD 5 ハッシュ関数を施して 1 2 8 ビットの要約値 AB 2 を求めると共に、そのヌクレオチドの配列の数 NA 2、及び先頭と末尾との 8 個ずつのヌクレオチドの配列 ST 2, SB 2 を求め、これらを内部の記憶装置の第 1 データファイルに記録する。テキストデータ TX 2 (図 9) に対する具体的な値は下記の通りである。

AB 2 = hex (1457b51222a83c3222e87cb4d4e63305) ... (2 5)

NA 2 = 2 0 4 8

ST 2 = A G C T T T T C, SB 2 = C G C G A A G G

次のステップ 1 2 4 において、情報処理装置は、試料 F の配列数 NA 2 と標準試料 E の配列数 NA 1 とが等しいかどうかを調べ、両者が異なっている場合には、ユーザはステップ 1 2 5 に移行して、別の DNA 情報を検索し、NA 2 と同じ配列数の DNA 情報をサーチする。本例では、ステップ 1 2 4 において、NA 2 = NA 1 であるため、動作はステップ 1 2 6 に移行して、試料 F の先頭と末尾との一部の配列 ST 2, SB 2 が、標準試料 E の配列 ST 1, SB 1 と等しいかどうか

か、更に試料Fの要約値AB2が標準試料Eの要約値AB1（ステップ121で一時ファイルに記録されている）と等しいかどうかを調べる。これらが共に等しい場合には、試料Fの配列と標準試料Eの配列とは非常に高い確率（ほぼ $1/2^{128} \approx 1/10^{38}$ 程度の確率）で一致しているとみなすことができる。従って、

5 ステップ127に移行して、コンピュータシステム2Bの情報処理装置は、その第1データファイルに「試料FのDNA構造は、標準試料EのDNA構造と同一である。」との情報を記録する。

但し、本例では、 $ST2 = ST1$ 、 $SB2 = SB1$ が成立するが、(11)式及び(25)式より $AB2 \neq AB1$ であるため、動作はステップ126からステップ128に移行して、その情報処理装置は、試料Fの先頭と末尾との一部の配列ST2、SB2が、標準試料Eを逆に並べた配列の一部の配列STR1、SBR1と等しいかどうか、更に試料Fの要約値AB2が標準試料Eを逆に並べた配列の要約値ABR1と等しいかどうかを調べる。これらが共に等しい場合には、

10 試料Fの配列と標準試料Eを逆に並べた配列とは非常に高い確率で一致しているとみなすことができる。従って、ステップ139に移行して、コンピュータシステム2Bの情報処理装置は、その第1データファイルに「試料FのDNA構造は、標準試料EのDNA構造に対して回文（palindrome）の関係にある。」との情報を記録する。

本例では、 $ST2 \neq STR2$ 、 $SB2 \neq SBR2$ 、かつ(12)式及び(2

20 5)式より $AB2 \neq ABR1$ であるため、動作はステップ128からステップ129に移行して、そのユーザは、通信ネットワーク1（インターネット）を介してDNA情報の供給者から上記の簡易データ、即ち標準試料Eのパリティ情報（ $B1(i) \sim B3(i)$ 、 $C1(j) \sim C3(j)$ ）（図8の情報）を購入し、購入した情報をコンピュータシステム2B（情報処理装置）内の記憶装置の第2

25 データファイルに記録する。

次に、図6のステップ130において、コンピュータシステム2Bの情報処理装置は、図9に示すように、試料FのテキストデータTX2を配列方向（ヌクレオチドの配列方向）にN行で、非配列方向にM列の16文字の長さの部分テキストデータTF(i, j)（ $i = 1 \sim N$ 、 $j = 1 \sim M$ ）に分割する。分割数N, M

は標準試料Eの分割数と同じであり、本例では、 $N=4$ 、 $M=32$ である。更に、情報処理装置は、図9の各部分テキストデータ $TF(i, j)$ を次のようにテキストデータを単にアスキーコードに変換する関数 $asc(TF(i, j))$ を用いて、 $128 (= 16 \cdot 8)$ ビットのバイナリーデータ（数値データ）よりなる部分データ $AF(i, j)$ に変換する。この場合にも、部分テキストデータ $TF(i, j)$ の文字列は反転してアスキーコード列に変換される。

$$AF(i, j) = asc(TF(i, j)) \quad \cdots (26)$$

この結果、図10に示す4行で、32列の部分データ $AF(i, j)$ が得られる。また、部分データ $AF(i, j)$ を連続して配列した集合体（数値データ）をバイナリーデータ $BN2$ とする。

次に、情報処理装置は、ステップ106の動作と同様にして、図10の各行（ $i=1 \sim 4$ ）の部分データ $AF(i, j)$ に対してガロア体 $GF(2^{128})$ 上で、非配列方向（ $j=1 \sim 32$ ）に対する和である第1パリティ(Parity) $B1F(i)$ 、 $\sum \alpha^{(j-1)} \cdot AF(i, j)$ である第2パリティ $B2F(i)$ 、及び $\sum \alpha^{2(j-1)} \cdot AF(i, j)$ である第3パリティ $B3F(i)$ を計算する。これらの非配列方向のパリティ $B1F(i) \sim B3F(i)$ （第1組のパリティ情報）は、

(15)式の生成元 α を用いて、かつ(14)式の既約多項式 $GF(X)$ を法として(19)式～(21)式と同様に、係数 i について1～4の範囲で計算される。

次に、その情報処理装置は、図10の各列（ $j=1 \sim 32$ ）の部分データ $AF(i, j)$ に対してガロア体 $GF(2^{128})$ 上で、配列方向（ $i=1 \sim 4$ ）に対する和である第1パリティ $C1F(j)$ 、 $\sum \alpha^{(i-1)} \cdot AF(i, j)$ である第2パリティ $C2F(j)$ 、及び $\sum \alpha^{2(i-1)} \cdot AF(i, j)$ である第3パリティ $C3F(j)$ を計算する。これらの配列方向のパリティ $C1(j) \sim C3(j)$ （第2組のパリティ情報）も、(15)式の生成元 α を用いて、かつ(14)式の既約多項式 $GF(X)$ を法として(22)式～(24)式と同様に、係数 j について1～32の範囲で計算される。

部分データ $AF(i, j)$ に対して実際にパリティ $B1F(i) \sim B3F(i)$ 、及びパリティ $C1F(j) \sim C3F(j)$ を計算した結果のベクトル表

示が、図10に16進数表示で示されている。

次に、ステップ131において、その情報処理装置は、供給者から購入した簡易データの2組のパリティ、即ち図8（標準試料E）の2組のパリティB1

(i) ~ B3 (i), C1 (j) ~ C3 (j) と、図10（試料F）の2組のパ

5 リティB1F (i) ~ B3F (i), C1F (j) ~ C3F (j) とを比較して、相違するパリティをサーチする。本例では、図8（標準試料F）に対して図10

（試料F）の $i = 1, 4$ の非配列方向のパリティB1F (1) ~ B3F (1),

B1F (4) ~ B3F (4) と、 $j = 16, 17$ の配列方向のパリティC1F

(16) ~ C3F (16), C1F (17) ~ C3F (17) とが異なっている。

10 なお、各行のパリティB1F (i) ~ B3F (i)、又は各列のパリティC1F (j) ~ C3F (j) において、1つでもパリティが異なっていれば、その行又は列のパリティが異なっているとみなすことができる。

従って、図10（試料F）の部分データAF (i, j) において、 $i = 1, 4$ の行と $j = 16, 17$ の列との交点に位置する4つの部分データAF (1, 1

15 6), AF (4, 16), AF (1, 17), AF (4, 17) が図8（標準試料E）と相違すると特定できる。また、これ以外の試料Fの部分データAF (i, j) は標準試料Eの部分データA (i, j) とほぼ同一であるとみなすことができる。

また、図11は、主に図10の試料Fのデータ中から図8と異なるパリティB

20 1F (1) ~ B3F (1), B1F (4) ~ B3F (4), C1F (16) ~ C

3F (16), C1F (17) ~ C3F (17) を取り出して表示したものである。

また、図11において図8と異なる部分データAF (1, 16), AF (4, 16), AF (1, 17), AF (4, 17) の位置に、復元すべきデータX1,

X2, Y1, Y2を表示している。この復元すべきデータX1, X2, Y1, Y

25 2はそれぞれ図8（標準試料E）の部分データA (1, 16), A (4, 16), A (1, 17), A (4, 17) である。

次のステップ132において、その情報処理装置は、図10の部分データAF

(i, j) 中で図8の部分データA (i, j) と相違する部分データ (AF (i', j')) とする) は、各行、又は各列に多くとも3つかどうかを調べる。これ

が成立する場合には、その部分データ $AF(i', j')$ に対応する標準試料 E の部分データ $A(i', j')$ は、ガロア体 $GF(2^{128})$ 上で連立方程式を解くことによって正確に求める（復元する）ことができる。本例では、それが成立する、即ち相違する変換データは、第 1 行、第 4 行に 2 つずつで、かつ第 1 6 列、

5 第 1 7 列に 2 つずつであるため、動作はステップ 1 3 3 に移行する。そして、その情報処理装置は、2 組の相違するパリティ、及び試料 F の相違する部分データ $AF(i', j')$ を用いて、図 1 2 のフローチャートに従って対応する標準試料 E の部分データ $A(i', j')$ (X_1, X_2, Y_1, Y_2) を復元する。図 1 2 の計算は、全てガロア体 $GF(2^{128})$ 上で実行される。

10 この場合、図 1 1 において、第 1 6 列の未知数 X_1, X_2 は 2 つであるため、第 1 6 列の 2 つのパリティ $C_1 F(16), C_2 F(16)$ と、対応する図 8 の 2 つのパリティ $C_1(16), C_2(16)$ と、未知数 X_1, X_2 に対応する試料 F の部分データ $AF(1, 16), AF(4, 16)$ とを用いて 2 元 1 次連立方程式を組み立てる。即ち、パリティ $C_1(16), C_1 F(16)$ に対する計

15 算式は図 1 2 のステップ 1 4 1 の (G 1) 式、(G 2) 式となる。また、(1 5) 式の生成元 α を用いて、パリティ $C_2(16), C_2 F(16)$ に対する計算式はステップ 1 4 2 の (G 3) 式、(G 4) 式となる。

次に、(G 1) 式から (G 2) 式を引き、(G 3) 式から (G 4) 式を引くことで、それぞれステップ 1 4 3 の (G 5) 式、(G 6) 式が得られる。(G 5)

20 式、(G 6) 式の右辺をそれぞれ $C_1 X, C_2 X$ とすることで、2 元 1 次連立方程式が得られる。そこで、これを解くことによって、未知数 X_1, X_2 はそれぞれステップ 1 4 4 の (G 7) 式で表すことができる。これを実際に解いた結果、 X_1, X_2 は次のようになる（図 1 1 参照）。なお、未知数が 3 個であれば、第 3 パリティ $C_3(16), C_3 F(16)$ などを用いて、3 元 1 次連立方程式

25 を解けばよく、未知数が 1 個であれば、例えば第 1 パリティ $C_1(16), C_1 F(16)$ などを用いるだけでよい。

$$X_1 = \text{hex}(43475447474347544347544354434154) \cdots (27)$$

$$X_2 = \text{hex}(47545447434147474341474754545441) \cdots (28)$$

更に、アスキーコード列を文字列に変換する関数 $\text{chr}()$ を用いて、この数

値データを文字列に変換すると次のようになる（図 1 1 参照）。この関数 $c h r$ （）は、上記の関数 $a s c$ （）と対称に、アスキーコード列を 1 バイト単位で最大桁のコードが末尾の文字となり、最小桁のコードが先頭の文字になるように反転して文字列に変換する。

$$\begin{aligned} 5 \quad c h r (X 1) &= TACTCTGCTGCGGTGC \\ &= T (1, 16) = T F (1, 16) \quad \cdots (29) \end{aligned}$$

$$c h r (X 2) = ATTTGGACGGACGTTG = T (4, 16) \quad \cdots (30)$$

これより、標準試料 E の部分テキストデータ $T (1, 16)$ と試料 F の部分テキストデータ $T F (1, 16)$ とは等しく、部分テキストデータ $T (4, 16)$ だけが部分テキストデータ $T F (4, 16)$ （図 9 参照）と異なることが分かる。

次に、図 1 1 において、第 1 7 列の未知数 $Y 1$, $Y 2$ についても、第 1 7 列の 2 つのパリティ $C 1 F (17)$, $C 2 F (17)$ と、対応する図 8 の 2 つのパリティ $C 1 (17)$, $C 2 (17)$ と、未知数 $Y 1$, $Y 2$ に対応する試料 F の部分データ $A F (1, 17)$, $A F (4, 17)$ とを用いて、図 1 2 のステップ 1 4 5 の (G 8) 式、(G 9) 式よりなる 2 元 1 次連立方程式が得られる。これを解くことによって、未知数 $Y 1$, $Y 2$ はそれぞれステップ 1 4 6 の (G 1 0) 式で表すことができる。これを実際に解いた結果、 $Y 1$, $Y 2$ は次のようになる（図 1 1 参照）。

$$Y 1 = \text{hex}(47544343415441544354474747474341) \quad \cdots (31)$$

$$20 \quad Y 2 = \text{hex}(41544343544754414743544741414754) \quad \cdots (32)$$

更に、この数値データ（アスキーコード列）を文字列に変換すると次のようになる（図 1 1 参照）。

$$c h r (Y 1) = \underline{ACGGGGTCT}ATACCTG = T (1, 17) \quad \cdots (33)$$

$$c h r (Y 2) = \underline{TGAAGTCGATGTCCTA}$$

$$25 \quad = T (4, 17) = T F (4, 17) \quad \cdots (34)$$

これより、標準試料 E の部分テキストデータ $T (4, 17)$ と試料 F の部分テキストデータ $T F (4, 17)$ とは等しく、部分テキストデータ $T (1, 17)$ だけが部分テキストデータ $T F (1, 17)$ （図 9 参照）と異なることが分かる。また、本例の方法によって未知数 $X 1$, $X 2$, $Y 1$, $Y 2$ 、即ち標準試料 E の部

分データ A (1, 16), A (4, 16), A (1, 17), A (4, 17) が正確に復元できていることが分かる。なお、部分データ A (1, 16), A (4, 17) は、それぞれ部分データ A F (1, 16), A F (4, 17) と同一であるため、復元されたデータとみなす必要はない。

- 5 次のステップ 134 において、その情報処理装置は、復元された部分データ A (i', j'), 即ち A (4, 16), A (1, 17) で、図 10 の試料 F のバイナリーデータ BN2 中の対応する部分データ A F (4, 16), A F (1, 17) を置き換えた後、この置き換えによって得られるバイナリーデータ BN2 をテキストデータ TX1' に逆変換する。更に情報処理装置は、そのテキストデータ TX1' より MD5 ハッシュ関数を用いて 128 ビットの要約値 AB1' を算出
- 10 し、この要約値 AB1' が標準試料 E の要約値 AB1 (ステップ 121 で一時ファイルに記録されている) と等しいかどうかを確認する。本例では、AB1' = AB1 が成立するが、例えば図 10 の試料 F の部分データ A F (i, j) 中の相違する部分のデータの状態によって、その相違がパリティ情報に反映されない
- 15 ような場合には、その相違する部分の位置がステップ 132 で正確に検出されない可能性がある。このような場合に、AB1' ≠ AB1 となったときには、ステップ 135 に移行すればよい。通常は、AB1' = AB1 が成立して、動作はステップ 138 に移行して、情報処理装置は、上記の第 1 データファイルに「試料 F の配列と標準試料 E の配列との内で相違する部分の位置 (i', j'), 及び
- 20 相違する部分テキストデータの対」の情報を記録する。本例では、位置 (i', j') として位置 (4, 16), (1, 17) が、相違する部分テキストデータの対として A (4, 16), A F (4, 16) 及び A (1, 17), A F (1, 17) が記録される。

- 一方、ステップ 132 において、相違する部分データ A F (i', j') の個
- 25 数が 4 個以上の行、又は列が存在する場合には、その行又は列での部分データの正確な復元は困難である。そこで、動作はステップ 135 に移行して、そのユーザはその DNA 情報の供給者から標準試料 E の完全データ、即ち図 8 のバイナリーデータ BN1 を通信ネットワーク 1 (インターネット) を介して購入し、コンピュータシステム 2B の情報処理装置は、そのバイナリーデータ BN1 を記憶装

置の第3データファイルに記録する。

次のステップ136において、その情報処理装置は、そのバイナリーデータBN1をテキストデータTX1'に逆変換し、そのテキストデータTX1'よりMD5ハッシュ関数を用いて128ビットの要約値AB1'を算出し、この要約値AB1'が標準試料Eの要約値AB1（ステップ121で一時ファイルに記録されている）と等しいかどうかを確認する。通常は、 $AB1' = AB1$ が成立するが、例えば通信エラー等によって送信されたバイナリーデータBN1の中にエラーが生じている場合には、 $AB1' \neq AB1$ となる。このときには、例えば供給者に完全データの再送信を要請する等の対処を行う。そして、ステップ136でAB1' = AB1が成立するときには、ステップ137に移行して情報処理装置は、標準試料EのバイナリーデータBN1中で、試料Fの相違している部分データAF(i', j')に対応する部分データA(i', j')を求める。その後、動作はステップ138に移行する。

なお、上記のステップ135では、ユーザはDNA情報の供給者から完全データ（バイナリーデータBN1）を購入しているが、別の方法として、ステップ131で特定された相違する部分データAF(i', j')に対応する標準試料Eの部分データA(i', j')のみを購入してもよい。これによって、通信コストを大幅に低減できる。

このように本例のビジネスモデルによれば、第1段階として標準試料Eのパリティ情報(B1(i) ~ B3(i), C1(j) ~ C3(j))を購入している。次に、このパリティ情報と試料Fのパリティ情報(B1F(i) ~ B3F(i), C1F(j) ~ C3F(j))とを比較して、相違する部分データAF(i, j)の個数が少ない場合には、対応する標準試料Eの部分データA(i, j)を復元することとして、相違する部分データの個数が多い場合に、完全データ、又は相違する部分データのみを購入している。従って、初めから膨大な完全データを購入する必要がなく、通信時間を短縮できると共に、情報処理コストを低減できる。

また、本例のパリティ情報を用いれば、SNP（一塩基変位多型：Single Nucleotide Polymorphism）のように所定の範囲内で1つのヌクレオチド（塩基）だ

けが異なっているような異常は、容易にその位置の検出、及び復元を行うことができる。

なお、上記の実施の形態では、DNA情報のユーザは、ステップ121において、コンテンツファイルより標準試料Eの配列ST1、SB1、要約値AB1、
5 及び配列STR1、SBR1、要約値ABR1を読み取って、ステップ122～128において、標準試料Eと試料Eとの同一性の判定を行って、両者が異なる場合に標準試料Eのパリティ情報（簡易データ）を購入している。しかしながら、標準試料Eと試料Fとは通常はいくらかは異なっていると考えられるため、このような要約値AB1等の読み取りから2つの試料の同一性の判定までの動作を省略して、すぐにステップ129に移行して、DNA情報の供給者から標準試料E
10 のパリティ情報（簡易データ）を購入するようにしてもよい。

なお、上記の実施の形態では、図8に示すように、第1組のパリティ情報（B1(i)～B3(i)）と第2組のパリティ情報（C1(j)～C3(j)）とは同じ個数である。しかしながら、図8の例のように部分データA(i, j)の
15 配列方向の個数（4個）が非配列方向の個数（32個）よりも少ない場合には、パリティ情報の全体の情報量を少なくするために、配列方向のパリティ情報として例えばC1(j)、C2(j)、又はC1(j)及びC2(j)のみを使用するというように、第2組のパリティ情報を第1組のパリティ情報よりも少なくしてもよい。このようにしても、標準試料Eと試料Fとの相違部は正確に検出でき
20 ると共に、SNPのような部分的に生じる相違部は正確に復元することができる。

また、図8（図7）の例のように部分データA(i, j)の配列方向の個数が非配列方向の個数よりも少ない場合には、その配列方向をディスプレイの横方向（水平方向）に合わせて表示し、非配列方向にスクロールすることによって、部分データA(i, j)及びパリティ情報を作業性の良好な状態で表示することが
25 できる。この場合でも、部分データA(i, j)をmビットとすると、その非配列方向の個数は、 $(2^m - 1) / 4$ 以下であることが望ましい。これによって、非配列方向のパリティ情報として異なる4個のパリティ情報を計算できるため、通常の部分的に生じる相違部は正確に復元することができる。

次に、上記の実施の形態では、ステップ105において、図7の標準試料Eの

テキストデータ TX 1 を図 8 の同じデータ量の部分データ A (i , j) の配列に変換し、この配列からパリティ情報を求めている。その代わりに、データ量を減少させるために、図 7 の標準試料 E のテキストデータ TX 1 を表 1 の変換テーブル（1つのヌクレオチドを 2 ビットのデータで表すテーブル）を用いてデータ量が 1 / 4 のバイナリーデータ（数値データ）に変換し、このバイナリーデータを
5 配列方向、及び非配列方向に分割して、部分データの配列を作成してもよい。

図 1 3 は、そのようにして得られたヌクレオチドの配列方向に 5 行で、非配列方向に 1 3 列の 6 4 ビット（8 バイト）ずつの部分データ B (i , j) （ i = 1 ~ 5 , j = 1 ~ 1 3 ）の配列を 1 6 進数表示で示し、この図 1 3 の各部分データ
10 B (i , j) は、それぞれ図 7 の標準試料 E 中の 3 2 個のヌクレオチドの配列に対応している。なお、この配列では、最後の部分データ B (5 , 1 3) に対応する図 7 の標準試料 E は存在しないため、その部分データ B (5 , 1 3) には hex (000...000) よりなるダミーデータが付加されている。

この場合には、ステップ 1 0 6 に対応して、図 1 3 の 6 4 ビットの各部分データ
15 B (i , j) をガロア体 GF (2 ⁶⁴) （ m = 6 4 ）上の元のベクトル表示とみなして、部分データ B (i , j) に対してガロア体 GF (2 ⁶⁴) 上の所定の演算を施す。ガロア体 GF (2 ⁶⁴) 上の既約多項式 GF (X) 及び生成元 α としては次の式を使用することができる。

$$GF(X) = 1 + X^5 + X^{23} + X^{43} + X^{64} \quad \cdots (35)$$

$$\alpha = X \quad \cdots (36)$$

また、ガロア体 GF (2 ⁶⁴) 上の既約多項式としては、例えば次の既約多項式 GF' (X) も使用でき、この既約多項式 GF' (X) に対する生成元としては次の α' などを使用できる。

$$GF'(X) = 1 + X^7 + X^{62} + X^{63} + X^{64} \quad \cdots (35A)$$

$$\alpha' = 1 + X \quad \cdots (36A)$$

そして、図 1 の情報処理装置 1 0 は、図 1 3 の各行 (i = 1 ~ 5) の部分データ B (i , j) に対してガロア体 GF (2 ⁶⁴) 上で、非配列方向 (j = 1 ~ 1 3) に対する和である第 1 パリティ (Parity) B 1 B (i) 、 $\sum \alpha^{(j-1)} \cdot B (i , j)$ である第 2 パリティ B 2 B (i) 、及び $\sum \alpha^{2(j-1)} \cdot B (i , j)$ である第

3パリティB3B(i)を計算する。これらの計算式は、(19)式～(21)式に対応しており、この第1組のパリティ情報(B1B(i)～B3B(i))はそれぞれ64ビットである。

更に、情報処理装置10は、図13の各列(j=1～13)の部分データB(i, j)に対してガロア体GF(2⁶⁴)上で、配列方向(i=1～5)に対する和である第1パリティC1B(j)、 $\sum \alpha^{(i-1)} \cdot B(i, j)$ である第2パリティC2B(j)、及び $\sum \alpha^{2(i-1)} \cdot B(i, j)$ である第3パリティC3B(j)を計算する。これらの計算式は、(22)式～(24)式に対応しており、この第2組のパリティ情報(C1B(j)～C3B(j))もそれぞれ64ビットである。

この場合、2組のパリティ情報(B1B(i)～B3B(i), C1B(j)～C3B(j))のデータ量は、図8のパリティ情報(B1(i)～B3(i), C1(j)～C3(j))に比べてほぼ1/4に少なくできる。従って、パリティ情報を通信回線を介して更に短時間で送信することができると共に、記録媒体に記録する場合にも、低容量の記録媒体を使用できる。

この場合には、ユーザ側で図9の試料Fのパリティ情報を計算する場合にも、同様にその試料Fのテキストデータを表1に従って1/4のデータ量の部分データの配列に変換した後、ガロア体GF(2⁶⁴)上の演算によって第1組及び第2組のパリティ情報を計算すればよい。この後の相違するデータの位置の特定、及び元のデータの復元は上記の実施の形態と同様に行うことができる。

なお、上記の実施の形態では、DNA又はRNAを構成するヌクレオチドは4種類であるため、テキストデータTX1を少ないデータ量のバイナリーデータに変換する際に、表1に示すように各ヌクレオチドを2ビットのデータで表している。これに対して、ヌクレオチド(又は塩基)を表すテキストデータとして、以下のような16種類の文字a～n(8ビットのアスキーデータ)が使用されることがある。

《表2》

- a アデニン (アデニンを含むヌクレオチドと同義、以下同様)
- c シトシン

	g	グアニン
	t	チミン
	u	ウラシル
	m	アデニン、又はシトシン
5	r	グアニン、又はアデニン
	w	アデニン、又はチミン（若しくはウラシル）
	s	グアニン、又はシトシン
	y	チミン（若しくはウラシル）、又はシトシン
	k	グアニン、又はチミン（若しくはウラシル）
10	v	アデニン、グアニン、又はシトシン
	h	アデニン、シトシン、又はチミン（若しくはウラシル）
	d	アデニン、グアニン、又はチミン（若しくはウラシル）
	b	グアニン、シトシン、又はチミン（若しくはウラシル）
	n	（アデニン、シトシン、グアニン、又はチミン（若しくはウラシル））
15		又は（不明若しくは他の塩基）。

この場合には、これら 16 種類の文字を互いに異なる 4 ビットのコードに変換し、このコードを用いてテキストデータを数値データ（バイナリーデータ）に変換してもよい。これによって、データ量を $1/2$ にすることができる。また、将来的にヌクレオチド（塩基）の種類が増加したような場合には、これらのヌクレ

20 オチドを 5 ビット、又は 6 ビットのデータで表現するようにしてもよい。

また、上記の実施の形態では、図 7 及び図 9 のヌクレオチドの配列を示すテキストデータよりハッシュ関数によって要約値を算出しているが、情報量としては、それらのテキストデータは例えば表 1 に従って変換したバイナリーデータ（数値データ）と等価である。従って、これらの変換後のバイナリーデータよりハッシュ

25 関数によってそれぞれ要約値を算出し、これらの算出結果同士を比較するようにしてもよい。そのバイナリーデータのデータ量はテキストデータに対して $1/4$ 程度であるため、要約値を算出する時間が短縮できる利点がある。

なお、上記の実施の形態では、DNA 又は RNA 中のヌクレオチドの配列（又は塩基の配列）の情報を処理対象としているが、本発明は、遺伝子を形成するヌ

クレオチドの配列の情報を処理する場合にも適用できることは言うまでもない。

次に、本発明の実施の形態の他の例につき説明する。本例は、タンパク質又はペプチドを構成するアミノ酸（生物学的物質）の配列情報を処理する場合に本発明を適用したものである。

- 5 本例でも基本的に図1のコンピュータシステム2Aを使用できるが、DNAのシーケンサー4の代わりに、タンパク質のアミノ酸の配列を決定する配列読み取り装置としてのタンパク質用のシーケンサー（protein Sequencer）が情報処理装置10に接続される点が異なっている。なお、その配列読み取り装置としては、アミノ酸の配列のデータベースも使用できる。本例でも、例えば新規の試料Gの
- 10 タンパク質のアミノ酸の配列をそのシーケンサーで解読した場合に、その配列を示すテキストデータ（TX3とする）が情報処理装置10に供給される。本例では一文字表記を採用するものとして、n個のアミノ酸の配列に対応するテキストデータ（アスキーコードとする）は、nバイトの長さである。本例では、その試料Gを大腸菌として、そのテキストデータTX3として、図14に示すように、
- 15 上記のウェブサイト1から入手した大腸菌の或るタンパク質の820個のアミノ酸の配列を示すテキストデータを使用する。

- 試料Gのアミノ酸配列は配列番号3に示されている。図14のテキストデータは、配列番号3の配列から数字データを除いて、その配列を一文字表記で表したものに相当する。また、図14においては、そのテキストデータが配列方向（ア
- 20 ミノ酸の配列方向）に4行で、その配列方向に直交する非配列方向に26列の8文字の長さの部分テキストデータTG（i，j）に分割されており、821番以上のアミノ酸を示すデータの位置にはダミーデータとして0が付加されている。

- 次に、情報処理装置10は、供給されたテキストデータTX3に上記のMD5ハッシュ関数を施して128ビットの要約値AB3を求めると共に、そのアミノ
- 25 酸の配列の数NA3、及び先頭と末尾との8個ずつのアミノ酸の配列ST3、SB3を求める。テキストデータTX3に対する具体的な値は下記の通りである。

AB3=hex(0f66dc2b3024a9739d0e912fde12b8ba) … (41)

NA3=820

ST3=MRVLKFGG, SB3=TL SWKLGV

次に、情報処理装置 10 は、テキストデータ TX 3 を逆方向に並べ替えたテキストデータ TXR 3 (=VGLKWS・・・FKLVRM) を求め、このテキストデータ TXR 3 の MD 5 ハッシュ関数による要約値 ABR 3、及びこのテキストデータ TXR 3 の先頭と末尾との 8 個ずつのアミノ酸の配列 STR 3, SBR 3 を求める。配列 STR 3, SBR 3 は、上記の配列 SB 3, ST 3 をそれぞれ逆方向に並べ替えることによって容易に求めることができる。これらの具体的な値は以下の通りである。テキストデータ TXR 3 の配列は、テキストデータ TX 3 の配列に対して回文 (palindrome) の関係にあるとすることができる。

ABR 3 = hex (e895f433ele77f84b3cadeeadla52380) ... (42)

STR 3 = VGLKWSLT, SBR 3 = GGFKLVRM

次に、情報処理装置 10 は、試料 G の名前の情報 (試料を特定する情報)、配列の数 NA 3、テキストデータ TX 3、配列 ST 3, SB 3、要約値 AB 3、逆方向の配列 STR 3, SBR 3、及び逆方向の要約値 ABR 3 を磁気ディスク装置 17 のマスターファイル 19 に記録する。この際に、マスターファイル 19 を複数のファイルとして、テキストデータ TX 3 と、それ以外のデータとを別のファイルに記録してもよい。続いて、情報処理装置 10 は、例えば図 7 と同様に図 14 に示すように、試料 G のテキストデータ TX 3 を配列方向 (アミノ酸の配列方向) に N 行で、その配列方向に直交する非配列方向に M 列の 8 文字 (64 ビット) の長さの部分テキストデータ TG (i, j) に分割する。N, M はそれぞれ 2 以上の任意の整数である。本例ではテキストデータ TX 3 に例えば 12 文字分のダミーデータ (本例では 0 であるが、それ以外に例えば文字 A など也可以使用できる) を付加して得られる 832 (=8・4・26) バイトのテキストデータ (これを TX 3' と呼ぶ) を作成し、テキストデータ TX 3' を N=4, M=26 で分割する。本例では、その 8 文字分の部分テキストデータ TG (i, j) を、次のようにテキストデータを単にアスキーコード (数値データ) に変換する関数 asc () を用いて、そのまま 64 ビットの部分データ AG (i, j) として扱う。

AG (i, j) = asc (TG (i, j)) ... (43)

なお、図 14 に TG (3, 11) の変換例で示すように、関数 asc (TG (i, j)) は、部分テキストデータ TG (i, j) の先頭の文字のコードが最

下位桁となり、末尾の文字のコードが最上位桁となるように反転して変換を行う。なお、この際に、各アミノ酸を6ビットのデータで表してもよいが、データ量は3/4程度になるだけであるため、本例では部分テキストデータ（アスキーコード列）をそのまま部分データ（数値データ）として扱う。

- 5 図15は、試料Gの部分データ $AG(i, j)$ の配列を示している。それに続いて図13の例と同様に、情報処理装置10は、その図15の4行で26列の64ビットの部分データ $AG(i, j)$ をガロア体 $GF(2^{64})$ ($m=64$) 上の元のベクトル表示とみなして、部分データ $AG(i, j)$ に対してガロア体 $GF(2^{64})$ 上の所定の演算を施す。ガロア体 $GF(2^{64})$ 上の既約多項式 $G(X)$ 及び生成元 α としては、(35)式（又は(35A)式）、及び(36)式（又は(36A)式）などを使用できる。

- 15 具体的に、情報処理装置10は、図15の各行 ($i=1\sim 4$) の部分データ $AG(i, j)$ に対してガロア体 $GF(2^{64})$ 上で、非配列方向 ($j=1\sim 26$) に対する和である第1パリティ (Parity) $B1G(i)$ 、 $\sum \alpha^{(j-1)} \cdot AG(i, j)$ である第2パリティ $B2G(i)$ 、及び $\sum \alpha^{2(j-1)} \cdot AG(i, j)$ である第3パリティ $B3G(i)$ を計算する。これらの計算式は、(19)式～(21)式に対応しており、この第1組のパリティ情報 ($B1G(i) \sim B3G(i)$) はそれぞれ64ビットである。

- 20 更に、情報処理装置10は、図15の各列 ($j=1\sim 26$) の部分データ $AG(i, j)$ に対してガロア体 $GF(2^{64})$ 上で、配列方向 ($i=1\sim 4$) に対する和である第1パリティ $C1G(j)$ 、 $\sum \alpha^{(i-1)} \cdot AG(i, j)$ である第2パリティ $C2G(j)$ 、及び $\sum \alpha^{2(i-1)} \cdot AG(i, j)$ である第3パリティ $C3G(j)$ を計算する。これらの計算式は、(22)式～(24)式に対応しており、この第2組のパリティ情報 ($C1G(j) \sim C3G(j)$) もそれぞれ64
25 ビットである。このように計算して得られたパリティ $B1G(i) \sim B3G(i)$ 、 $C1G(j) \sim C3G(j)$ が、図15に16進数表示で示されている。

この例においては、各行、各列で3つのパリティを計算しているが、アミノ酸の配列はヌクレオチドの配列に比べるとデータ量がかなり少ないため、実用的には、パリティ情報は各行（非配列方向）、及び各列（配列方向）においてそれぞれ

れ1つ（例えばB 2 G（i）とC 2 G（j））としてもよい。図15の例において、パリティB 2 G（i），C 2 G（j）のみを使用するものとする、パリティはそれぞれ64ビット（8バイト）であるため、全部のパリティのデータ量は、240（＝8・30）バイトとなる。従って、全部のパリティのデータ量は、
5 体の元のテキストデータTX 3（820バイト）に対してほぼ1／3に減少している。

次に、情報処理装置10は、試料Gの名前の情報、配列の数NA 3、テキストデータTX 3、要約値AB 3、ABR 3、及びパリティ情報を磁気ディスク装置17のワーキングファイル20に記録する。この際に、ワーキングファイル20
10 を複数のファイルとしてもよい。その後、情報処理装置10は、試料Gの名前の情報、配列の数NA 3、配列ST 3、SB 3、要約値AB 3、逆方向の配列STR 3、SBR 3、及び逆方向の要約値ABR 3を磁気ディスク装置17のコンテンツファイル21に記録する。更に、情報処理装置10は、コンテンツファイル21中の情報を通信ネットワーク1を介してコンテンツのプロバイダ3に送信する。
15 これによって、コンテンツファイル21中の情報はプロバイダ3のサーバ内の閲覧可能なコンテンツファイル31に記録されて、第3者がインターネットを介して自由に閲覧できるようになる。この結果、第3者は、公開されている試料Gの配列の数NA 3、及び要約値AB 3（又は必要に応じてABR 3）を自分の保有するアミノ酸の配列の配列数、及び要約値と比較することによって、その試
20 料Gが自分にとって新規かどうかを判定できる。また、ユーザは、その試料Gの配列情報を複数の供給者から誤って重複して購入することを回避することができる。

その後、コンピュータシステム2Aの所有者（アミノ酸情報の供給者）は、ユーザから購入要求が来るのを待つ状態となる。そして、ユーザから試料Gに対する簡易データの要求があったときには、情報処理装置10は、磁気ディスク装置
25 17のワーキングファイル20の中の試料Gのパリティ情報（例えばその中のB 2 G（i），C 2 G（j））を例えば電子メールの添付ファイルとしてそのユーザに送信する。パリティの情報を購入したユーザは、試料Gと同じ種類の自分で解読した試料のアミノ酸の配列のパリティと、その購入したパリティとを比較す

ることによって、相違する部分の検出及び復元を或る程度行うことができる。

一方、ユーザから完全データの要求があったときには、情報処理装置 10 は、ワーキングファイル 20 中のテキストデータ TX 3 を ZIP ファイル等の形式で圧縮し、この圧縮されたデータを例えば電子メールの添付ファイルとしてそのユーザに送信する。この際に必要に応じて、ハッシュ関数による要約値 AB 3 を同時に送信してもよい。本例によれば、簡易データ（パリティ情報）はデータ量を少なくできるために短時間で送信することができる。

更に、そのアミノ酸の配列情報の供給者は、ワーキングファイル 20 に記録した情報、即ち試料 G の名前の情報、配列の数 NA 3、テキストデータ TX 3、要約値 AB 3、ABR 3、及びパリティ情報を CD-R/RW ドライブ 15 を介して CD-R 16 に記録してもよい。この CD-R 16 から、更に多数の CD-ROM を作製してもよく、これらの記録媒体が郵送等によってユーザに販売される。

なお、上記の実施の形態では、生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが m ビット（ m は 16 以上の整数）の部分データに分割し、複数行の前記部分データに各行毎に前記非配列方向にガロア体 $GF(2^m)$ 上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体 $GF(2^m)$ 上の第 2 の演算を施して第 2 組のパリティ情報を求めている。

これに関して、 m が 16 より小さい場合には、各部分データは、例えば 1～7 個程度のヌクレオチド、又は 1～2 個程度のアミノ酸の配列に対応するため、1 つの生物学的物質の配列に対して計算すべきパリティ情報の数が多くなり過ぎて好ましくない。また、最近のコンピュータの計算能力を活かしきれないという不都合もある。特に、コンピュータの処理単位が 64 ビットの倍数である場合には、パリティ情報の計算効率を高めるために、その m の値も 64、128、192、256 などの 64 の倍数とすることが望ましい。

また、例えば図 13 の実施の形態では、 m ビットを超える素数を $P (> 2^m)$ とすると、この素数 P を法とするガロア体 $GF(P)$ を用いてパリティ情報を計

算することも可能である。しかしながら、このガロア体 $GF(P)$ を用いた場合には、 m ビットの部分データに演算を施して得られる個々のパリティ情報が m ビットを超える場合があるため、パリティ情報がガロア対 $GF(2^m)$ を用いた場合に比べて $(m+1)/m$ 倍程度に長くなるという不都合がある。即ち、ガロア

5 体 $GF(P)$ には演算が容易であるという利点があるが、ガロア体 $GF(2^m)$ を用いた場合には、個々のパリティ情報を m ビットにできるため、パリティ情報を簡潔に記録できる利点がある。

2^m ($m \geq 16$) を超える、又は 2^m より僅かに小さい程度の大きい値の素数 P を求めるには、例えば ミラー・ラビン (Miller-Rabin) の素数判定法 (例えば

10 M. O. Rabin: "Probabilistic algorithms for testing primality", Journal of Number Theory, 12, pp. 128-138 (1980) 参照) を使用することができる。また、生成元を求めるために大きい数の因数分解を行うためには、例えば 2 次ふるい法 (quadratic sieve) (例えば C. Pomerance: "Factoring", In Cryptology and Computational Number Theory, pp. 27-47, American Mathematical Society (1990)

15 参照) を使用することができる。また、実際に大きい数の素数 P の決定、及び大きい数の因数分解を行うためには、上記の「UBASIC」中の組み込み関数「ECM」を使用してもよい。

そして、素数 P を用いるガロア体 $GF(P)$ は、 $(0, 1, \dots, P-1)$ の P 個の元より構成され、加減乗除の演算は P を法として実行される。これを上記の

20 実施の形態に適用する場合には、 m ビットの部分データ ($A(i, j)$ 等) をそのガロア体 $GF(P)$ の何れかの元に対応させればよい。そして、パリティ情報を容易に算出するためには、ガロア体 $GF(P)$ の生成元 δ を使用するとよい。ガロア体 $GF(P)$ 上の生成元 δ は、 $k = P-1$ とおくと、 P を法として、次の関係を満たす。

25
$$\delta^k = 1 \pmod{P} \quad \dots (A2)$$

$$\delta^{k'} \neq 1 \pmod{P} \quad (1 \leq k' < k) \quad \dots (A3)$$

そこで、素数 p_1, p_2, \dots, p_r 及び整数 n_1, n_2, \dots, n_r を用いて、 k が次のように因数分解できるものとする。

$$k = P-1 = p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r} \quad \dots (A4)$$

このとき、生成元 δ とは、 P を法として、 δ の $(p_1^{n_1-1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r})$ 乗、 $(p_1^{n_1} \cdot p_2^{n_2-1} \cdot \dots \cdot p_r^{n_r})$ 乗、 \dots 、 $(p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r-1})$ 乗が何れも 1 とならないものであればよい。

また、ガロア体 $GF(P)$ 上の任意の 0 以外の元 ε についても (A 2) 式が成立するため、 $k (=P-1)$ を用いて ε の逆元 ε^{-1} は次のように計算することも可能である。

$$\varepsilon^{-1} = \varepsilon^{k-1} \pmod{P} \quad \dots (A 5)$$

従って、例えば或る部分データを ε で除算する場合には、その部分データに ε^{k-1} を乗算すればよい

そして、ガロア体 $GF(P)$ を用いる場合にも、図 3～図 6 の DNA 情報の供給者の動作、及び DNA 情報のユーザの動作は、ガロア $GF(2^m)$ 上の演算をガロア体 $GF(P)$ 上の演算に変えるだけで全て同様に適用することができる。

具体的に図 1 3 (例えば標準試料とする。) の実施の形態において、ガロア体 $GF(P)$ 上でパリティ情報を計算する場合には、部分データ $B(i, j)$ は 64 ビットの全ての値を取る可能性があるため、素数 P としては 2^{64} ($m=64$) を超える範囲で $65 (=m+1)$ ビット以下の素数 (例えば最も小さい素数) を選択すればよい。このような素数 P は 16 進数表示で次のようになる。また、この素数 P に対する生成元 δ としては、例えば次の値を使用できる。なお、この場合、例えば 3 も生成元であるため、生成元として 2 のべき乗以外の数を使用したいときには、生成元 δ として例えば 3 を使用してもよい。

$$P = \text{hex}(10000000000000000d) \quad \dots (A 6)$$

$$\delta = 2 \quad \dots (A 7)$$

この条件で、図 3 のステップ 106 に対応して、図 1 3 の 64 ビットの各部分データ $B(i, j)$ をガロア体 $GF(P)$ 上の元のとみなして、部分データ $B(i, j)$ に対して生成元 δ を用いてガロア体 $GF(P)$ 上の所定の演算を施す。一例として、図 1 3 の各行 ($i=1 \sim 5$) の部分データ $B(i, j)$ に対してガロア体 $GF(P)$ 上で、非配列方向 ($j=1 \sim 13$) に対する和である第 1 パリティ $B_1 B(i)'$ 、 $\sum \delta^{(j-1)} \cdot B(i, j)$ である第 2 パリティ $B_2 B(i)'$ 、及び $\sum \delta^{2(j-1)} \cdot B(i, j)$ である第 3 パリティ $B_3 B(i)'$ を計算す

る。これらの計算式は、(19)式～(21)式に対応しており、この第1組の
パリティ情報 ($B1B(i)' \sim B3B(i)'$) はそれぞれ最大で $(64 +$
1) ビットである。

更に、図13の各列 ($j = 1 \sim 13$) の部分データ $B(i, j)$ に対してガロ
5 ア体 $GF(P)$ 上で、配列方向 ($i = 1 \sim 5$) に対する和である第1パリティ $C1B(j)'$ 、 $\sum \delta^{(i-1)} \cdot B(i, j)$ である第2パリティ $C2B(j)'$ 、及
び $\sum \delta^{2(i-1)} \cdot B(i, j)$ である第3パリティ $C3B(j)'$ を計算する。こ
これらの計算式は、(22)式～(24)式に対応しており、この第2組のパリティ
10 ィ情報 ($C1B(j)' \sim C3B(j)'$) もそれぞれ最大で $(64 + 1)$ ビッ
トである。

同様に図6のステップ130に対応して、比較対象の試料の部分データの配列
に対してガロア体 $GF(P)$ 上で非配列方向の第1組のパリティ情報、及び配
列方向の第2組のパリティ情報を計算する。そして、本例でも図6のステップ1
31～138に対応して、2つの試料のパリティ情報の比較、相違する部分デー
15 タの特定、及び相違する部分データの復元を行うことができる。

なお、部分データが128ビット ($m = 128$) で全ての値を取る場合に、ガ
ロア体 $GF(P)$ 上でパリティ情報を計算する場合には、 2^{128} を超えて129
ビット以下の素数 P 、及びこれに対応する生成元 δ の一例としては、次の値を使
用することができる。

$$20 \quad P = 2^{128} + 51 \quad \cdots (A8)$$

$$\delta = 2 \quad \cdots (A9)$$

次に、図7、図8の実施の形態のように部分データ $A(i, j)$ がテキストデ
ータそのものである場合には、 m ビットの部分データ $A(i, j)$ の最大値 Nm
 $a x$ は $(2^m - 1)$ よりも小さくなる。具体的に図7のテキストデータとしてA
25 S C I Iコードを使用する場合、アルファベットのA S C I Iコードはhex(41)
～hex(7a) であるため、最大値 $Nm a x$ は次のようになる。

$$N m a x = \text{hex}(7a7a7a7a \cdots 7a7a7a) < 2^m - 1 \quad \cdots (A10)$$

この場合には、ガロア体 $GF(P)$ を構成するための素数 P を次のように、部
分データ $A(i, j)$ の最大値 $Nm a x$ より大きく、且つ m ビットの数として選

$$2^m \geq P \geq N_{\max} \quad \dots (A11)$$

P=hex(7a7a7a7a7a7a7a7a7a7a7a7a7a7a7a7f) ... (A 1 2)

$$\delta = 5 \quad \dots \quad (\text{A } 1 \text{ } 3)$$

10 単であると共に、ガロア体 $GF(2^m)$ を用いたときと同じ長さの簡潔なパリティ情報を得ることができる。

$$15 \quad N_{\max} = \text{hex}(7a7a7a7a7a7a7a7a) < 2^m - 1 \quad \cdots (A14)$$

具体的に、 $m=64$ で(A14)式が成立する場合には、(A11)式を満たす範囲で最も小さい素数 N 及びこれに対する生成元 δ の一例は次のようになる。

P=hex (7a7a7a7a7a7a7ad5) ... (A 1 5)

$$\delta = 2 \quad \dots \quad (\text{A } 16)$$

25 の配列方向の個数を、その部分データの非配列方向の個数よりも少なくして、配列方向（第2組）のパリティ情報の個数を、非配列方向（第1組）のパリティ情報の個数よりも少なくしてもよい。これによって、パリティ情報をディスプレイに表示し易いと共に、パリティ情報の情報量が必要以上に大きくなるのを避けることができる。この場合、生成元 δ を用いて非配列方向で4個までの相違部を正

確に復元するためには、非配列方向の部分データの個数を $(P - 1) / 4$ 以下にすればよい。

なお、上記の実施の形態では、各部分データに所定の係数を乗じてパリティ情報を求めているが、パリティ情報としては、例えば BCH 符号 (Bose-Chaudhari-Hocquenghem Codes) (例えば J. L. Massey: "Shift Register Synthesis and BCH Decoding", IEEE Trans., IT-15, pp. 122-127 (1969) 参照) を用いてもよい。但し、従来の BCH 符号は小さい値 m' ($m \leq 16$ 程度) のガロア体 $GF(2^{m'})$ で考えられているため、これを本発明で使用されている大きい値 m ($m \geq 16$) のガロア体 $GF(2^m)$ 上で構築し直す必要がある。

ここで、上記の実施の形態で使用するハッシュ関数に関して説明する。通常、暗号理論で使われるハッシュ関数は、テキストデータ中のスペースコード及び改行コード等も全て演算処理対象としているが、ヌクレオチド及びアミノ酸の配列情報については見やすくするために、例えば配列番号 1 ~ 3 で示すように、途中にスペースコード、順序を示す数字コード、及び改行コードを挿入する場合がある。そこで、ヌクレオチドやアミノ酸などの生物学的物質の配列情報を演算処理対象とするハッシュ関数においては、テキストデータ中の所定コードとしてのスペースコード、数字コード、及び改行コードを無視する機能を付加することが望ましい。また、隣接する文字を " - " (ハイフン) で分けることも考えられるが、この場合には、更に " - " 記号も無視する必要がある。また、例えばファイルの最後に「データの終わりを示すコード」が付加されるような場合には、そのコードも無視するようにしてもよい。

また、例えばヌクレオチドの配列が、通常は小文字で表されるような場合には、ハッシュ関数に、選択的に大文字を小文字に変換して要約値を計算する機能を持たせるようにしてもよい。逆に、例えばアミノ酸の配列が、通常は大文字で表されるような場合には、ハッシュ関数に、選択的に小文字を大文字に変換して要約値を計算する機能を持たせるようにしてもよい。

更に、原ファイルを複数の分割ファイルに分割する際には、複数の分割ファイルの順序等を示すデータ (以下、「コメントデータ」と言う) を各分割ファイルに付加することが望ましいことがある。このように分割ファイル、又は 1 つの原

ファイルにコメントデータを付加する場合にも、コメントデータはハッシュ関数で無視する必要がある。そのため、例えばコメントデータは所定の開始記号（例えば ／＊ ）及び終了記号（例えば ＊／ ）の間に記録し、ハッシュ関数で処理する際に開始記号から終了記号までのデータは無視するようにすればよい。

- 5 また、上記の実施の形態では、例えば生物のDNAのヌクレオチドの配列（又はタンパク質のアミノ酸の配列）内の先頭の一部、及び末尾の一部の配列、並びにその配列のテキストデータの要約値をインターネット上で公開することがある。この場合には、その公開されている一部の配列と、その要約値とからそのテキストデータの10 内容が推定される可能性もある。これを回避するために、そのテキストデータをハッシュ関数で処理する際に、その公開されている配列を除いた部分についてのみ、そのハッシュ関数を施して要約値を求めるようにしてもよい。

- なお、上記の実施形態の図5のステップ124において、標準試料Eの配列の数NA1と試料Fの配列の数NA2とが k （ k は1以上の整数）だけ異なる場合で、例えば $NA2 = NA1 + k$ のときには、試料FのテキストデータTX2から15 k 個のヌクレオチド分のデータを取り除いてから、相違部を求めるためにステップ126に移行してもよい。また、 $NA2 = NA1 - k$ のときには、そのテキストデータTX2に k 個のヌクレオチド分のダミーデータを付加してから、ステップ126に移行してもよい。これによって、試料Fの配列の過剰部又は欠損部の位置を特定することができる。このときに、 $k = 1$ であるときには、テキストデータTX220 中でデータを除去又は付加する位置は、全長の $1/2$ の位置、その前部又は後部の $1/2$ （全体の $1/4$ ）の位置、更にその全部又は後部の $1/2$ （全体の $1/8$ ）の位置、…としていくことで、最も短時間に過剰部又は欠損部の位置を特定できる。

- なお、本発明は上述の実施の形態に限定されず、本発明の要旨を逸脱しない範囲で25 種々の構成を取り得ることは勿論である。

産業上の利用の可能性

本発明によれば、核酸や遺伝子中のヌクレオチド、又は又はタンパク質やペプチド中のアミノ酸などの生物学的物質の配列情報を、それらの配列を示すテキス

トデータよりも少ないデータ量のパリティ情報として近似的に記録することができる。従って、そのパリティ情報は、低容量の記録媒体にも記録できると共に、通信回線を介して短時間に送信することが可能となる。また、ガロア体 $GF(2^m)$ 上の演算を行うことによって、個々のパリティ情報を部分データと同じ m ビットの情報量で簡潔に記録できる利点がある。

一方、そのパリティ情報を求める場合に、素数 P のガロア体 $GF(P)$ 上の演算を行うことによって、パリティ情報の情報量は $(m+1)/m$ 倍程度に多くなるが、演算を単純化することができる。

特に、部分データの最大値 N_{max} が $(2^m - 1)$ よりも小さい値であり、その素数 P を、 $(2^m > P > N_{max})$ の関係を満たすように選択できる場合には、演算を単純化した上で、各パリティ情報を部分データと同じ m ビットの情報量で簡潔に記録できる利点がある。

また、2つの生物学的物質の配列のパリティ情報を比較することによって、2つの配列間の相違部の位置を少ないデータ量で容易に特定（検出）できると共に、必要に応じてその相違部の情報を復元することができる。従って、例えば SNP（一塩基変位多型：Single Nucleotide Polymorphism）を少ないデータ量で容易に発見することができる。

また、本発明によれば、ヌクレオチド又はアミノ酸などの生物学的物質の配列情報を近似する情報（パリティ情報）を少ないデータ量でユーザに供給できるビジネスモデルを提供することができる。この場合に、更に数学的な要約値を用いることによって、ユーザが提供された配列情報と情報供給者が保持している配列情報との同一性の確認などを容易に行うことができる。また、同一の複数の配列情報を誤って購入することも防止できる。

請 求 の 範 囲

1. 生物学的物質の配列情報の記録方法であって、

5 前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを
所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方
向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが m ビッ
ト (m は16以上の整数)の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体 $GF(2^m)$ 上
の第1の演算を施して第1組のパリティ情報を求めると共に、

10 複数列の前記部分データに各列毎に前記配列方向にガロア体 $GF(2^m)$ 上の
第2の演算を施して第2組のパリティ情報を求め、

前記第1組及び第2組のパリティ情報で前記生物学的物質の配列を表すことを
特徴とする生物学的物質の配列情報の記録方法。

2. 請求の範囲1に記載の記録方法であって、

15 前記ガロア体 $GF(2^m)$ 上の生成元を α としたとき、

前記第1組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記
非配列方向に順次 $\alpha^{s \cdot p}$, $\alpha^{s \cdot (p+1)}$, $\alpha^{s \cdot (p+2)}$, ..., $\alpha^{s \cdot (p+d \cdot p)}$ (s は0以上の整数、
 p は0以上の整数、 $d \cdot p$ は1以上の整数)を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

20 前記第2組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記
配列方向に順次 $\alpha^{t \cdot q}$, $\alpha^{t \cdot (q+1)}$, $\alpha^{t \cdot (q+2)}$, ..., $\alpha^{t \cdot (q+d \cdot q)}$ (t は0以上の整数、
 q は0以上の整数、 $d \cdot q$ は1以上の整数)を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴と
する生物学的物質の配列情報の記録方法。

25 3. 請求の範囲2に記載の記録方法であって、

前記部分データの前記配列方向の個数は、前記部分データの前記非配列方向の
個数よりも少なく、

前記第2組のパリティ情報の個数は、前記第1組のパリティ情報の個数よりも
少ないことを特徴とする生物学的物質の配列情報の記録方法。

4. 請求の範囲 3 に記載の記録方法であって、

前記部分データの前記非配列方向の個数は、 $(2^m - 1) / 4$ 以下であることを特徴とする生物学的物質の配列情報の記録方法。

5. 請求の範囲 2 に記載の記録方法であって、

5 前記整数 s 及び t は 0 であることを特徴とする生物学的物質の配列情報の記録方法。

6. 請求の範囲 2 に記載の記録方法であって、

前記整数 s 及び t は 1 であることを特徴とする生物学的物質の配列情報の記録方法。

10 7. 請求の範囲 2 に記載の記録方法であって、

前記第 1 組のパリティ情報は、前記複数行の各行毎に前記整数 s について互いに異なる複数の値で求めた複数の和を含み、

前記第 2 組のパリティ情報は、前記複数列の各列毎に前記整数 t について互いに異なる複数の値で求めた複数の和を含むことを特徴とする生物学的物質の配列
15 情報の記録方法。

8. 請求の範囲 1 に記載の記録方法であって、

前記生物学的物質の配列中の各生物学的物質をそれぞれ 6 ビット以下の数値データで表して得られる数値データを前記演算の対象とすることを特徴とする生物学的物質の配列情報の記録方法。

20 9. 請求の範囲 1 に記載の記録方法であって、

前記ガロア体 $GF(2^m)$ を規定する整数 m は 64 の倍数であることを特徴とする生物学的物質の配列情報の記録方法。

10. 請求の範囲 1 に記載の記録方法であって、

前記生物学的物質の配列を基準配列として、該基準配列の前記 2 組のパリティ
25 情報に対応させて、検査対象の生物学的物質の配列について前記 2 組のパリティ情報を求め、

前記 4 組のパリティ情報より前記基準配列に対する前記検査対象の生物学的物質の配列の相違部を求めることを特徴とする生物学的物質の配列情報の記録方法。

11. 請求の範囲 1 に記載の記録方法であって、

前記生物学的物質は、DNA、RNA、又は遺伝子の少なくとも一部を構成するヌクレオチドであることを特徴とする生物学的物質の配列情報の記録方法。

1 2. 請求の範囲 1 に記載の記録方法であって、

5 前記生物学的物質は、一つのタンパク質の少なくとも一部を構成するアミノ酸であることを特徴とする生物学的物質の配列情報の記録方法。

1 3. 生物学的物質の配列情報の記録装置であって、

前記生物学的物質の配列情報を読み取る配列読み取り装置と、

10 前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが m ビット (m は 16 以上の整数) の部分データに分割するデータ配列手段と、

15 複数行の前記部分データに各行毎に前記非配列方向にガロア体 $GF(2^m)$ 上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体 $GF(2^m)$ 上の第 2 の演算を施して第 2 組のパリティ情報を求める演算手段と、

前記第 1 組及び第 2 組のパリティ情報を記録媒体に記録する記録手段とを有することを特徴とする生物学的物質の配列情報の記録装置。

1 4. 請求の範囲 1 3 に記載の記録装置であって、

前記ガロア体 $GF(2^m)$ 上の生成元を α としたとき、

20 前記第 1 組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次 $\alpha^{s \cdot p}$, $\alpha^{s \cdot (p+1)}$, $\alpha^{s \cdot (p+2)}$, ..., $\alpha^{s \cdot (p+d \cdot p)}$ (s は 0 以上の整数、 p は 0 以上の整数、 $d \cdot p$ は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

25 前記第 2 組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次 $\alpha^{t \cdot q}$, $\alpha^{t \cdot (q+1)}$, $\alpha^{t \cdot (q+2)}$, ..., $\alpha^{t \cdot (q+d \cdot q)}$ (t は 0 以上の整数、 q は 0 以上の整数、 $d \cdot q$ は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の記録装置。

1 5. 生物学的物質の配列情報を記録したコンピュータ読み取り可能な記録媒体

であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが m ビット（ m は16以上の整数）の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体 $GF(2^m)$ 上の第1の演算を施して第1組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体 $GF(2^m)$ 上の第2の演算を施して第2組のパリティ情報を求め、

- 10 前記生物学的物質の配列に関する情報が、前記第1組及び第2組のパリティ情報として記録されたことを特徴とするコンピュータ読み取り可能な記録媒体。

16. 請求の範囲15に記載の記録媒体であって、

- 15 前記生物学的物質の配列に対応する前記テキストデータ、又は該テキストデータに対応する前記数値データの40ビット以上の長さの数学的な要約値が更に前記記録媒体に記録されたことを特徴とするコンピュータ読み取り可能な記録媒体。

17. 生物学的物質の配列情報の供給方法であって、

- 20 前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを保持する供給者が、前記テキストデータ、又はこれに対応する前記数値データを第1ファイルに記録して保持する第1ステップと、

前記第1ファイルに記録されている前記テキストデータ、又は該テキストデータに対応する前記数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが m ビット（ m は16以上の整数）の部分データに分割し、

- 25 複数行の前記部分データに各行毎に前記非配列方向にガロア体 $GF(2^m)$ 上の第1の演算を施して第1組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体 $GF(2^m)$ 上の第2の演算を施して第2組のパリティ情報を求める第2ステップと、

前記供給者が、前記第1組及び第2組のパリティ情報を第2ファイルに記録し

て保持する第 3 ステップと、

前記生物学的物質の配列情報のユーザが、通信回線を介して前記供給者より前記第 2 ファイルに記録されている前記 2 組のパリティ情報を受け取る第 4 ステップと

- 5 を有することを特徴とする生物学的物質の配列情報の供給方法。

18. 請求の範囲 17 に記載の供給方法であって、

前記ユーザが、前記 2 組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内の前記供給者の生物学的物質の配列情報との相違部を特定する第 5 ステップと、

- 10 該相違部の配列の復元ができない場合に、前記ユーザが前記通信回線を介して前記供給者より前記第 1 ファイルに記録されている前記テキストデータ、又は前記数値データの内の前記配列の復元ができない部分の配列情報を受け取る第 6 ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

- 15 19. 請求の範囲 18 に記載の供給方法であって、

前記ガロア体 $GF(2^m)$ 上の生成元を α としたとき、

前記第 1 組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次 $\alpha^{s \cdot p}$, $\alpha^{s \cdot (p+1)}$, $\alpha^{s \cdot (p+2)}$, ..., $\alpha^{s \cdot (p+d \cdot p)}$ (s は 0 以上の整数、 p は 0 以上の整数、 $d \cdot p$ は 1 以上の整数) を乗算した後、

- 20 該演算で得られた複数の積について各行毎に求められた和を含み、

前記第 2 組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次 $\alpha^{t \cdot q}$, $\alpha^{t \cdot (q+1)}$, $\alpha^{t \cdot (q+2)}$, ..., $\alpha^{t \cdot (q+d \cdot q)}$ (t は 0 以上の整数、 q は 0 以上の整数、 $d \cdot q$ は 1 以上の整数) を乗算した後、

- 25 該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の供給方法。

20. 請求の範囲 18 に記載の供給方法であって、

前記供給者は、前記生物学的物質の配列の長さの情報、及び前記配列を表すテキストデータ又は前記数値データの数学的な要約値の情報を前記通信回線を介して閲覧可能な状態にしておき、

前記ユーザは、前記第4ステップの前に前記通信回線を介して前記配列の長さの情報及び前記数学的な要約値の情報を閲覧することを特徴とする生物学的物質の配列情報の供給方法。

2 1. 生物学的物質の配列情報の記録方法であって、

- 5 前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが m ビット (m は16以上の整数)の部分データに分割し、

- 10 前記部分データの最大値を N_{max} 、この最大値 N_{max} よりも大きい素数を P として、

複数行の前記部分データに各行毎に前記非配列方向にガロア体 $GF(P)$ 上の第1の演算を施して第1組のパリティ情報を求めると共に、

複数列の前記部分データに各列毎に前記配列方向にガロア体 $GF(P)$ 上の第2の演算を施して第2組のパリティ情報を求め、

- 15 前記第1組及び第2組のパリティ情報で前記生物学的物質の配列を表すことを特徴とする生物学的物質の配列情報の記録方法。

2 2. 請求の範囲2 1に記載の記録方法であって、

前記部分データの最大値 N_{max} は $(2^m - 1)$ よりも小さい値であり、
前記素数 P は、

- 20 $2^m > P > N_{max}$

の関係を満たすことを特徴とする生物学的物質の配列情報の記録方法。

2 3. 請求の範囲2 2に記載の記録方法であって、

前記ガロア体 $GF(P)$ 上の生成元を δ としたとき、

- 25 前記第1組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次 $\delta^{s \cdot p}$, $\delta^{s \cdot (p+1)}$, $\delta^{s \cdot (p+2)}$, ..., $\delta^{s \cdot (p+d \cdot p)}$ (s は0以上の整数、 p は0以上の整数、 $d \cdot p$ は1以上の整数)を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第2組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次 $\delta^{t \cdot q}$, $\delta^{t \cdot (q+1)}$, $\delta^{t \cdot (q+2)}$, ..., $\delta^{t \cdot (q+d \cdot q)}$ (t は0以上の整数、

q は 0 以上の整数、d q は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の記録方法。

2 4. 生物学的物質の配列情報の供給方法であって、

- 5 前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを保持する供給者が、前記テキストデータ、又はこれに対応する前記数値データを第 1 ファイルに記録して保持する第 1 ステップと、

- 10 前記第 1 ファイルに記録されている前記テキストデータ、又は該テキストデータに対応する前記数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが m ビット (m は 1 6 以上の整数) の部分データに分割し、

前記部分データの最大値を N_{max} 、この最大値 N_{max} よりも大きい素数を P として、

- 15 複数行の前記部分データに各行毎に前記非配列方向にガロア体 $GF(P)$ 上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体 $GF(P)$ 上の第 2 の演算を施して第 2 組のパリティ情報を求める第 2 ステップと、

- 20 前記供給者が、前記第 1 組及び第 2 組のパリティ情報を第 2 ファイルに記録して保持する第 3 ステップと、

前記生物学的物質の配列情報のユーザが、通信回線を介して前記供給者より前記第 2 ファイルに記録されている前記 2 組のパリティ情報を受け取る第 4 ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

- 25 2 5. 請求の範囲 2 4 に記載の供給方法であって、

前記ユーザが、前記 2 組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内の前記供給者の生物学的物質の配列情報との相違部を特定する第 5 ステップと、

該相違部の配列の復元ができない場合に、前記ユーザが前記通信回線を介して

前記供給者より前記第 1 ファイルに記録されている前記テキストデータ、又は前記数値データの内の前記配列の復元ができない部分の配列情報を受け取る第 6 ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

要 約 書

ヌクレオチド又はアミノ酸などの生物学的物質の配列情報を少ないデータ量で近似的に記録する記録方法及び装置である。標準試料EのDNAを構成する一列のヌクレオチドの配列を示すテキストデータを所定の変換規則に従ってバイナリーデータに変換し、このバイナリーデータを複数行で複数列の m ビット ($m \geq 16$) の部分データ ($A(i, j)$) に分割する。各行の部分データ ($A(i, j)$) を非配列方向にガロア体 $GF(2^m)$ 上で演算して第1組のパリティ ($B1(i) \sim B3(i)$) を求め、各列の部分データ ($A(i, j)$) を配列方向にガロア体 $GF(2^m)$ 上で演算して第2組のパリティ ($C1(j) \sim C3(j)$) を求め、これらのパリティ情報によってヌクレオチドの配列を近似的に表す。